Dispensa di Statistica di Base (Prof. Carmine Fago)

Prefazione

Nella mia carriera di studente prima e docente poi, nei vari corsi di studi, mi sono frequentemente imbattuto nella solita e ricorrente domanda che quasi tutti gli studenti si pongono: “Cos’è la statistica?” e soprattutto “A cosa serve la statistica?”

La risposta, di natura didascalica, sembrerebbe semplice e recuperabile in qualsivoglia dizionario: si definisce analisi statistica un’indagine tendente alla valutazione quantitativa dei comportamenti ossia l'insieme dei metodi d'indagine tendenti al raggiungimento di valutazioni quantitative del comportamento umano o animale.

Possiamo accontentarci di tale fredda definizione? Proviamo a capire meglio cos’è la statistica…

Questa risulta essere una lente di ingrandimento sul nostro mondo…

Noi uomini siamo abituati a significare i fenomeni attraverso i nostri cinque sensi, fermandoci spesso, solo allo strato più superficiale del quanto ed utilizzando quale codice comunicativo la parola, ossia quell’insieme di lettere che vanno a racchiudere la semantica dell’evento osservato. La statistica, dal canto suo, attraverso i simboli “numerici”, penetra la realtà, trascendendo l’individuale percezione condizionata dalla soggettività, rendendola immediatamente fruibile ed obbiettivamente interpretabile nei suoi aspetti più nascosti. In sostanza questa assurge ad essere l’applicazione matematica del pensiero stesso.

Il nostro studio può essere paragonato ad un’avventura che trova il suo inizio nell’esplorazione di una realtà apparentemente alternativa; realtà che, svestita dei suoi diversi strati, sembra essere sempre mutabile e mutata, per poi finire nel comprendere che non è cambiata nei suoi massimi sistemi, ma molto più semplicemente, diversamente palesata agli occhi di noi uomini che, acquisito un nuovo strumento, abbiamo avuto la possibilità di scrutarla nella sua totalità. Tale strumento è appunto la statistica… quella conoscenza che ci permette di leggere la realtà in maniera più precisa, obbiettiva ed attenta, attraverso i numeri.

Per gli studenti della facoltà di Scienze Internazionali, essa rappresenta un mezzo essenziale per prendere decisioni informate, valutare l’impatto delle politiche pubbliche e studiare le dinamiche socioeconomiche globali.

Indice

Introduzione

Capitolo I – Introduzione alla statistica

- Cos'è la statistica?
- Cosa studia la statistica?
- Obiettivi di questo corso
- Il metodo statistico: funzione descrittiva e funzione inferenziale
- La variabile: definizione e sua operazionalizzazione
- Le scale di misurazione
- Presentazione dei principali indici statistici (tendenza centrale e variabilità)
- Rappresentazioni grafico-matematiche per sintetizzare i dati (tabelle di frequenza/tabelle a doppia entrata)

Capitolo II – Misure di tendenza centrale

- La moda
- La Mediana
- La media aritmetica

Capitolo III – Gli indici di posizione

- I quantili

Capitolo IV – La variabilità: gli indici di variabilità o dispersione

- Range o Campo di variazione
- Differenza interquartile o scarto interquartile
- Scostamento semplice medio o Scarto medio assoluto
- Devianza
- Varianza
- Deviazione standard o Scarto quadratico medio o Scarto tipo
- Coefficiente di variazione

Capitolo V – Dai punti Z agli indici di forma

- I punti standard
- La curva normale o Gaussiana
- La curva normale standardizzata
- Gli indici di forma

Capitolo VI – Accenni all'intervallo di fiducia o confidenza

Capitolo VII – La statistica bivariata

- Tabella di contingenza
- Frequenze relative condizionate
- Indipendenza di due caratteri
- Connessione ed Indice Chi quadrato
- La dipendenza in media

- Coefficiente di correlazione lineare di Pearson
- Coefficiente di determinazione ed alienazione
- Retta di regressione lineare secondo il metodo dei minimi quadrati

Capitolo VIII – Gini

- Indice di eterogeneità
- Indice di concentrazione

Capitolo I – Introduzione alla statistica

Cos’è la statistica?

Sai ched'è la statistica? È na' cosa che serve pe fà un conto in generale de la gente che nasce, che sta male, che more, che va in carcere e che spósa. Ma pè me la statistica curiosa è dove c'entra la percentuale, pè via che, lì, la media è sempre eguale puro co' la persona bisognosa. Me spiego: da li conti che se fanno seconno le statistiche d'adesso risurta che te tocca un pollo all'anno: e, se nun entra nelle spese tue, t'entra ne la statistica lo stesso perch'è c'è un antro che ne magna due. Er compagno scompagno: lo che conosco bene l'idee tue so' certo che quer pollo che te magni, se vengo giù, sarà diviso in due: mezzo a te, mezzo a me... Semo compagni. No, no - rispose er Gatto senza core - io non divido gnente co' nessuno: fo er socialista quanno sto a diggiuno, ma quanno magno so' conservatore. (Trilussa)

Cos’è e cosa studia la statistica?

Cerchiamo di capirlo… Cos’è la realtà? Tutto ciò che noi sentiamo, vediamo, tocchiamo, altro non è che un insieme di impulsi elettrici interpretati dal nostro cervello e di conseguenza ognuno di noi ha una sua personale percezione del reale del fenomeno osservato. Esistono quindi diverse realtà a seconda della persona che osserva… La statistica cerca di fornire un’interpretazione del tutto, quanto più univoca ed obiettiva possibile.

Definizione di Statistica

Scienza che studia i fenomeni collettivi, attraverso metodi matematici- Scienza che ha come scopo lo studio quantitativo e qualitativo di un particolare fenomeno collettivo in condizioni di incertezza o non determinismo, cioè di non completa conoscenza di esso o di una sua parte.- Scienza che studia i fenomeni collettivi, sia naturali che sociali, attraverso metodi matematici, fondati soprattutto sulle tecniche di campionamento e sul calcolo delle probabilità, allo scopo di tracciare modelli esplicativi e di formulare previsioni … ognuno si diverte a dare la propria versione…

Obiettivi di questo corso

Saper riconoscere, leggere, interpretare, confrontare e magari creare uno studio statistico che rappresenti un dato fenomeno:

	M (DS)	Range	1	2	3	4	5
1. Età	20.66 (2.70)	19-53	1				
2. Esami sostenuti	7.10 (5.13)	0-30	-	1			
3. Autoefficacia accademica	3.92 (0.63)	1-5	0.00	0.02	1		
4. Adattamento accademico	3.70 (0.85)	1-5	-0.07	-0.05	0.34**	1	
5. Locus of Control interno	26.37 (3.53)	6-36	-0.01	0.10*	0.51**	0.29**	1

Note. *$p < 0.05$; **$p < 0.01$.

Il metodo statistico

FUNZIONE DESCRITTIVA

(Analisi univariata/bivariata/multivariata)

Inerente al metodo attraverso il quale è possibile riassumere tutte quelle informazioni riguardanti il fenomeno oggetto di studio, rendendole più facilmente fruibili ed utilizzabili.

18 21 23 21 26 27 30 30 28 28

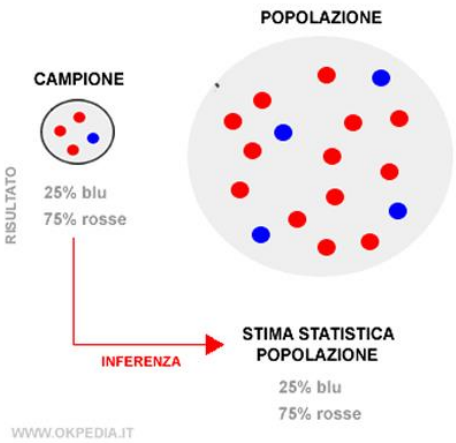
27 27 26 25 25 24 18 18 19 19

30 30 28 28 27 26 24 24 18 22

X_i	fi	Fi
18	4	4
19	2	6
21	2	8
22	1	9
23	1	10
24	3	13
25	2	15
26	3	18
27	4	22
28	4	26
30	4	30

Funzione Inferenziale (Induttiva)

Partendo dal particolare (Campione), permette di inferire o generalizzare le informazioni oggetto di analisi alla popolazione di riferimento, generando leggi e/o proprietà d'ordine generale.



La variabile: definizione e sua operazionalizzazione

La variabile è quel particolare aspetto delle unità statistiche che costituisce l'oggetto dell'indagine, avente quali proprietà l'esaustività e l'esclusività.

Tale può essere:

- Qualitativa/Categoriale (nominali/sconnesse – ordinali)

- Quantitativa (discrete – continue)

Ed anche (nella statistica inferenziale)

- Indipendente (normalmente indicata con la lettera X)

- Dipendente (normalmente indicata con la lettera Y)

- Interveniente (di disturbo)

Cos'è l'operazionalizzazione:

L'operazionalizzazione di un costrutto è il processo di individuazione dei comportamenti osservabili che ci permettono di rilevare la presenza di tale caratteristica in un soggetto. Devono essere comportamenti osservabili da chiunque.

Cos'è un costrutto: Nel campo della ricerca, un costrutto si riferisce ad un concetto o a un'idea astratta che viene esaminata o studiata. Rappresenta un'entità teorica o ipotetica che non può essere osservata o misurata direttamente.

Come giungere alla variabile, partendo da un costrutto:

Esempio:

- Costrutto oggetto di studio: aggressività (non direttamente osservabile)

- Individuazione di quelle manifestazioni comportamentali rappresentative del costrutto (direttamente osservabili): scatti d'ira (indicatore)

- Definizione operativa della variabile: decido come misurare l'indicatore

- Ho la mia variabile, misurando la quale, posso inferire rispetto al costrutto originario

La variabile

	VARIABILE			
	CATEGORIALE		QUANTITATIVA	
LIVELLO	NOMINALE	ORDINALE	DISCRETA	CONTINUA
CARATTERISTICHE	CATEGORIE DISTINTE	CATEGORIE ORDINATE	NON FRAZIONABILI	FRAZIONABILI
OPERAZIONI	= ≠	= ≠ < >	= ≠ < > + -	= ≠ < > + - X /

Scale di misurazione

SCALE DI MISURAZIONE SECONDO LA CLASSIFICAZIONE DI STEVENS			
TIPO DI VARIABILE	**PROCESSO UTILIZZATO PER OTTENERE IL DATO**	**MODALITA' D'ESPRESSIONE DEL DATO**	**TIPO DI SCALA DI MISURAZIONE**
QUALITATIVA/CATEGORIALE	CLASSIFICAZIONE QUALITATIVA	DESCRIZIONE X CATEGORIA (MASCHIO/FEMMINA)	SCALA NOMINALE
	CLASSIFICAZIONE ORDINALE	ATTRIBUTO ORDINALE (ADOLESCENTE/ADULTO/ANZIANO)	SCALA ORDINALE
QUANTITATIVA	CONTEGGIO	NUMERO NATURALE (1,2,3…)	SCALA AD INTERVALLI EQUIVALENTI (ZERO CONVENZIONALE)
	MISURA	NUMERO RAZIONALE (2.51, 15.18…)	SCALA A RAPPORTI EQUIVALENTI (ZERO ASSOLUTO)

Indici di tendenza centrale e variabilità in relazione alle scale di misurazione

		SCALE DI MISURAZIONE			
	INDICI STATISTICI	**SCALA NOMINALE (VARIBILI NOMINALI)**	**SCALA ORDINALE (VARIABILI ORDINALI)**	**SCALA AD INTERVALLI EQUIVALENTI (VARIABILI QUANTITATIVE)**	**SCALA A RAPPORTI EQUIVALENTI (VARIABILI QUANTITATIVE)**
INDICI DI TENDENZA CENTRALE	**MODA**	SI	SI	SI	SI
	MEDIANA		SI	SI	SI
	MEDIA			SI	SI
INDICI DI VARIABILITA'	**SCARTO INTERQUARTILE (DIFF. INTERQUARTILE)**		SI	SI	SI
	CAMPO DI VARIAZIONE (RANGE - GAMMA)			SI	SI
	SCOSTAMENTO SEMPLICE MEDIO			SI	SI
	VARIANZA			SI	SI
	DEVIAZIONE STANDARD (SCARTO TIPO - SCARTO QUADRATICO MEDIO)			SI	SI
	COEFFICIENTE DI VARIAZIONE			SI	SI

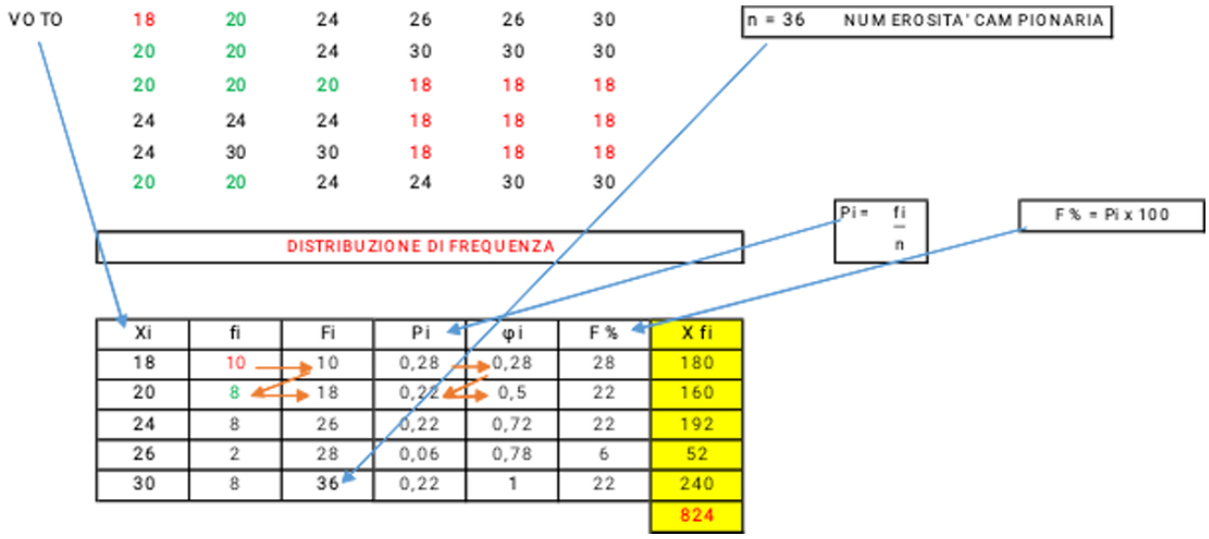
Rappresentazioni grafico-matematiche per sintetizzare i dati (tabelle di frequenza e tabelle a doppia entrata)

<u>Distribuzione semplice</u>

- 18 23 19 24 30 28 27 19 20 21 18 19 19 20 21 23 24 27 28 30

Nota: in tale distribuzione, prima di procedere con l'analisi si rende necessario ordinare i valori dal minore al maggiore.

Distribuzione di frequenza



Xi	fi	Fi	Pi	φi	F %	X fi
18	10	10	0,28	0,28	28	180
20	8	18	0,22	0,5	22	160
24	8	26	0,22	0,72	22	192
26	2	28	0,06	0,78	6	52
30	8	36	0,22	1	22	240
						824

Distribuzioni di frequenza in classi

X	fi	Fi	Pi	F %
18--20	18	18	0,18	18
21--23	24	42	0,24	42
24--26	16	58	0,16	58
27--28	24	82	0,24	82
29--30	18	**100**	0,18	100

Tabella a doppia entrata

		X = GENERE		
		M	**F**	**Ri**
Y = VOTO	**21**	5	12	17
	24	13	6	19
	26	4	2	6
	30	8	12	20
	Ci	30	32	**62**

Y	**Fi**	**Fcum**
21	17	17
24	19	**36**
26	6	**42**
30	20	**62**

X	**Fi**	**Fcum**
M	30	30
F	32	**62**

Fo = frequenze osservate (frequenze congiunte)
ampiezza della tabella a doppia entrata = 4
Ri = marginale di riga
Ci = marginale di colonna
Fe = valori che un ricercatore si aspetterebbe di trovare all'interno delle celle, avendo quei determinati marginali e quella determinata numerosità campionaria per i quali le due variabili risulterebbero indipendenti

Capitolo II – Misure di tendenza centrale

Gli indici di tendenza centrale sono degli indicatori sintetici che, in un unico valore (o pochi valori), riassumono l'intera distribuzione dei dati, nel loro aspetto più tipico, dando, in tal maniera, un'idea dell'ordine di grandezza del fenomeno studiato. A seguire andremo a descrivere più nel dettaglio i tre principali indicatori sintetici: la moda, la mediana e la media aritmetica.

La Moda (Mo)

Tra gli indici di tendenza centrale, la moda o norma (che solitamente si indica con Mo) è il più semplice da determinare e può essere calcolata per tutti i tipi variabile e quindi su ogni scala di misurazione.

Definizione: per le distribuzioni semplici e di frequenza, la moda è la modalità della variabile a cui è associata frequenza assoluta (relativa o (relativa) percentuale) più elevata. Per le distribuzioni di frequenza in classi, la moda è rappresentata dalla classe che presenta densità di frequenza più elevata.

A seconda del numero di valori/modalità modali, una distribuzione può essere:

- Unimodale (un solo valore/modalità modale)
- Bimodale (due valori/modalità modali)
- Multimodale (tre o più valori/modalità modali)
- Amodale (non vi sono valori/modalità modali in quanto tutte le frequenze assolute/densità sono tra loro uguali)

Esempio distribuzione semplice

2 4 6 8 10 12 12 14 14 14 → Mo = 14 (fi=3 maggiore rispetto alle altre)

Esempio distribuzione di frequenza

X	fi	Fi
18	18	18
20	6	24
24	3	27
26	3	30
30	6	**36**

Mo = 18 (fi=18 maggiore rispetto alle altre)

Esempio distribuzione di frequenza

X	fi	Fi	i	Densità
18--20	18	18	3	6
21--23	12	30	3	4
24--26	5	35	3	1,666667
27--28	6	41	2	3
29--30	13	**54**	2	**6,5**

La classe modale è la (29--30)

La Mediana (Me)

La mediana è quell'indice di tendenza centrale (posizione) che "taglia" la distribuzione in due parti uguali, dividendo i dati esattamente al 50%. Calcolabile a partire dalla scala ordinale, risulta essere la misura di tendenza centrale robusta per quelle distribuzioni eccessivamente asimmetriche. Essendo un indice di posizione, per calcolarlo bisognerà dapprima calcolarne la posizione per poi verificare il valore della variabile corrispondente alla posizione trovata.

Per le distribuzioni semplici distinguiamo due casi:

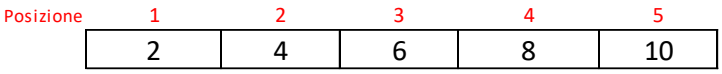
1. Se n è dispari, la posizione mediana è uguale a:

$$\frac{n+1}{2}.$$

2. Se n è pari, la posizione mediana sarà compresa tra:

$$\frac{n}{2} \quad \text{e} \quad \frac{n}{2}+1$$

Esempio

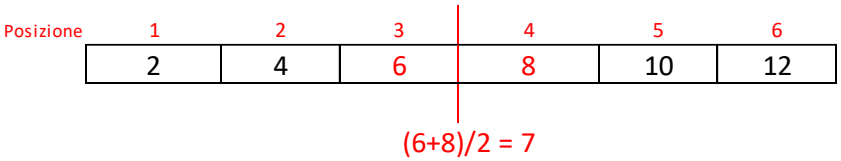


n = 5 DISPARI

POS mediana = (n+1)/2

POS mediana = (5+1)/2 = 3

Mediana = 6 (in quanto il valore 6 si trova in terza posizione



n = 6 PARI

POS mediana = N/2 < Pos mediana < (N/2) + 1

3 < pos med < 4

Per le distribuzioni di frequenza, avendo normalmente una numerosità campionaria più importante, potremo calcolare la posizione mediana semplicemente dividendo per 2 la stessa numerosità campionaria (n/2). Una volta calcolata tale posizione, si andrà a verificare la corrispondenza tra questa e la modalità della variabile relativa al primo valore della frequenza cumulata (Fi) contenente la stessa posizione.

Esempio

X	fi	Fi
18	10	10
20	11	21
24	5	26
26	7	33
30	13	**46**

Pos med = n/2 = 46/2 = 23

23

N = numerosità campionaria

Me = 24

Per le distribuzioni di frequenza in classi, oltre a calcolare la posizione mediana e verificare la classe corrispondente attraverso il primo valore della Fi che contiene la stessa posizione, andremo ad individuare il valore preciso della mediana all'interno della classe sopramenzionata applicando la formula dell'interpolazione.

Formula dell'interpolazione

$$\text{lim.reale inf.cl.med} + \frac{(\text{pos.med} - \text{Fi.cl.inf.cl.med}) \times \text{i.cl.med}}{\text{fi.cl.med}}$$

Esempio

X	fi	Fi
18--20	8	8
21--23	6	14
24--27	6	20
28--30	10	30

moda = (28--30)
mediana = 24,16

Pos.med = N/2 = 15

INTERPOLAZIONE X CALCOLO MEDIANA

lim.reale inf.cl.med + (pos.med - Fi.cl.inf.cl.med) x i.cl.med
fi.cl.med

lim reale inf = lim tabulato inf - 0,5 → 23,5
lim reale sup = lim tabulato sup + 0,5 → 27,5
i.cl.med = lim reale sup - lim reale inf 27,5-23,5 = 4

mediana = 23,5 + (15 - 14) x 4 / 6 = 23,5 + 0,66 = 24,16

La Media aritmetica

L'ultimo indice di posizione che analizziamo è la media aritmetica, che può essere calcolata solo per i caratteri quantitativi, quindi su scale ad intervalli equivalenti e rapporti equivalenti. Diviene misura di tendenza centrale robusta (rappresentativa) per quelle distribuzioni non eccessivamente asimmetriche, in quanto condizionata dai così detti valori estremi altrimenti detti valori outliers.

La media aritmetica di n valori numerici {x1, . . . , xi , . . . , xn}, è data dalla seguente formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Proprietà della media da approfondire

Proprietà formali

Internalità

Omogeneità

Associativa

Linearità

Annullamento

Ottimizzazione

Proprietà di minimo della media aritmetica

Esempi di calcolo

X	fi	Fi	Xi fi
4	8	8	32
6	6	14	36
8	6	20	48
10	10	**30**	100
			216

Media= ∑ Xi fi / n

Media: 216/30 = 7,2

Xi = valore della variabile

Xifi = prodotto tra il valore della variabile e la frequenza assoluta

X	fi	Fi	Xc	Xc fi
18--20	8	8	19	152
21--23	6	14	22	132
24--27	6	20	25,5	153
28--30	10	**30**	29	290
				727

Media= ∑ Xc fi / n

Media: 727/30 = 24,2333

Xc = valore centrale di classe

Xcfi = prodotto tra il valore centrale di classe e la frequenza assoluta

Capitolo III – Gli indici di posizione: i quantili

I quantili sono quei valori che dividono la distribuzione in parti uguali. Nel caso si abbiano classi di valori si usa talvolta "supporre" che i valori siano distribuiti in modo uniforme all'interno di ciascuna classe, in modo da calcolare il quantile (per interpolazione) su una funzione continua.

I quantili di ordini "semplici", ad esempio quelli espressi come frazioni (cioè quando α è un numero razionale), vengono anche chiamati con altri nomi. I quantili di ordini 1/n, 2/n, ..., (n-1)/n dividono la popolazione in n parti ugualmente popolate; il quantile di ordine α=m/n è detto **m-esimo n-ile**.

- La **mediana** è il quantile di ordine 1/2.
- I **quartili** sono i quantili di ordini 1/4, 2/4 e 3/4.

Altri particolari quantili sono:

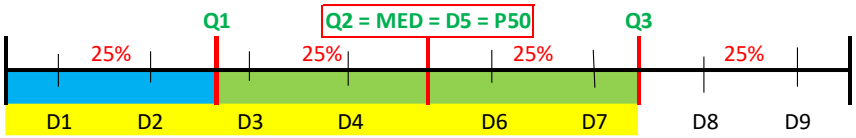
- I **decili**, di ordine m/10, dividono la popolazione in 10 parti uguali.
- I **centili o percentili**, di ordine m/100, dividono la popolazione in 100 parti uguali.

Notare che, dividendo la distribuzione in parti uguali, avremo alcuni quantili che, pur denominati in diversa maniera, risultano essere ed indicare il medesimo valore:

- Mediana = 2° quartile (Q2) = 5° decile (D5) = 50° percentile (P50)
- 1° quartile (Q1) = 25° percentile (P25)
- 3° quartile (Q3) = 75° percentile (P75)
- … E così via.

Il giusto procedimento per calcolare il quantile risulta essere:

- Individuare dapprima la posizione quantilica mediante la formula [(n+1)J]/K per le distribuzioni semplici e di frequenza o (nJ)/K per le distribuzioni di frequenza in classi, ove n = numerosità campionaria; K = le parti nelle quali vogliamo dividere la distribuzione, J = il quantile specifico che vogliamo calcolare.
- Calcolata la posizione, si andrà a verificare il valore della variabile oggetto d'analisi corrispondente alla posizione stessa.
- In caso di distribuzioni di frequenza in classi, dopo aver individuato la classe quantilica, si andrà a calcolare il valore preciso del quantile mediante la formula dell'interpolazione.



POSquantilica = $\frac{(n+1) \times J}{K}$ (J → il quantile specifico; K → numero di parti nelle quali dividiamo la distribuzione)

Esempi

Distribuzione semplice

2 4 6 8 10 12

POS Q3 = [(6+1)3]/4 = 5,25

(Altri esempi di calcolo pos.)
POS Q1 = [(6+1)1]/4 = 1,75
POS Q2 = [(6+1)2]/4 = 3,5
POS D4 = [(6+1)4]/10 = 2,8

Dopo aver trovato la posizione, calcoliamo il valore del Q3 mediante la formula dell'interpolazione lineare…

FORMULA DELL'INTERPOLAZIONE LINEARE VAL MIN + (VAL MAX - VAL MIN)DEC. POS

… seguendo il seguente procedimento: partendo dalla posizione del Q3 = 5,25, possiamo vedere che il valore del nostro Q3 cade tra la quinta e la sesta posizione ossia tra i valori 10 e 12. Considerando tale coppia di valori, applichiamo la sopra menzionata formula considerando:

- VAL MIN = 10
- VAL MAX = 12
- DEC POS (parte decimale della posizione) = 0,25

10 + (12 - 10)0,25 = 10 + 2X0,25 = 10 + 0,5 = 10,5

Distribuzione di frequenza

X	fi	Fi
18	18	18
20	6	24
24	3	27
26	3	30
30	6	**36**

$$\text{Pos Q3} = \frac{(36 + 1)3}{4} \longrightarrow 27{,}75$$

$$\text{Pos Q1} = \frac{(36 + 1)1}{4} \longrightarrow 9{,}25$$

Una volta trovata la posizione del Q3 = 27,75, andremo ad individuare il primo valore della Fi che contiene la posizione; il Q3 corrisponderà al corrispondente valore della variabile. Nel caso specifico la posizione del Q3 = 27,75 è compreso in Fi = 30 che corrisponde ad X = 26. Il nostro Q3 sarà quindi uguale a 26.

Distribuzione di frequenza in classi

X	fi	Fi
18--20	18	18
21--23	12	30
24--26	5	35
27--28	6	41
29--30	13	**54**

$$\text{Pos q} = \frac{NJ}{K}$$

Q1 = (54 x 1)/4 = 13,5

La classe quantilica corrispondente è la (18—20) essendo Fi = 18 il primo valore contenente 13,5.

Applicando la formula dell'interpolazione, andremo a calcolare il valore preciso del Q1 all'interno della classe individuata

$$\text{lim.reale inf.cl.quant} + \frac{(\text{pos.quant} - \text{Fi.cl.inf.cl.quant}) \times \text{i.cl.quant}}{\text{fi.cl.quant}}$$

$$Q1 = [17,5 + (13,5 - 0)3]/18 = 19,75$$

Capitolo IV – La variabilità: gli indici di variabilità o dispersione

Oltre a voler stabilire l'ordine di grandezza del fenomeno studiato, si può essere interessati a valutare l'attitudine del carattere a variare, ossia ad assumere valori distinti. In questo caso si parla di studio della variabilità (o mutabilità) del fenomeno. Gli indici che misurano l'attitudine di un carattere ad assumere modalità diverse (a variare) sono: gli indici di variabilità per le variabili quantitative e gli indici di eterogeneità per le variabili qualitative. Nel nostro corso presentiamo solo alcuni degli indici di variabilità esistenti in letteratura e tralasciamo completamente gli indici di eterogeneità.

Il più semplice indice di variabilità è il range o campo di variazione: $R = x(n) - x(1)$. Se la distanza tra il più grande e il più piccolo valore osservato è piccola, significa che il carattere assume valori molto prossimi tra loro e quindi varia poco. D'altra parte non si può affermare con certezza il contrario qualora il range sia grande, poiché può accadere che la maggior parte dei dati sia concentrata in un certo intervallo di valori (tali osservazioni variano poco tra loro) e poche osservazioni (anche solo una) siano molto diverse, in questo caso il range risulterebbe alto ma la variabilità dei dati (escludendo queste poche osservazioni) non è elevata. Per ovviare questo problema, si è deciso di costruire indici di variabilità che escludono i valori più piccoli e più elevati della serie di dati. Un indice di questo tipo è la differenza (e lo scarto) interquartile, che essendo dato dalla differenza (o semidifferenza) tra il terzo e il primo quartile non è influenzato dalle osservazioni inferiori al primo quartile o superiori al terzo. Tuttavia sia la formula di calcolo del range che quella della differenza interquartile dipende solo da due dei valori osservati, non tiene conto di tutti i dati.

Indici di variabilità costruiti sulla base dell'intera serie di dati sono lo scostamento semplice medio, la varianza e lo scarto quadratico medio. Proprio perché dipendono da tutte le osservazioni, la varianza e lo scarto quadratico medio (come pure la media) sono molto influenzati dai valori anomali (outliers), ossia osservazioni insolitamente alte o basse. In altri termini, i valori della varianza e dello scarto quadratico medio (come pure quello della media) cambiano notevolmente se si includono o meno i valori anomali presenti nella serie di dati. Per questa ragione, talvolta è preferibile calcolare media e varianza dopo aver eliminato dalla serie di dati i valori anomali stessi. Oppure si fa riferimento alla mediana e alla differenza interquartile, poiché non dipendono dalla presenza delle osservazioni anomale, nel senso che non cambiano di valore se queste osservazioni vengono escluse o meno dalla serie di dati.

Gli Indici di Variabilità o Dispersione

In statistica, un indice di dispersione è un indice che descrive sinteticamente la variabilità di una distribuzione statistica quantitativa. In modo particolare misura quanto i valori presenti nella distribuzione distano da un valore centrale scelto come riferimento. Il valore centrale è solitamente un indice di posizione. I valori centrali più usati nelle applicazioni sono media e mediana.

Gli indici di variabilità che prendiamo in considerazione sono:

- Range o Campo di variazione
- Differenza interquartile o scarto interquartile
- Scostamento semplice medio o Scarto medio assoluto
- Devianza
- Varianza
- Deviazione standard o Scarto quadratico medio o Scarto tipo
- Coefficiente di variazione

Range o Campo di variazione

Questo altro non è che la differenza tra il valore più grande e quello di piccolo della distribuzione

$$\textbf{range o campo di variazione dei dati: } R = x_{(n)} - x_{(1)}$$

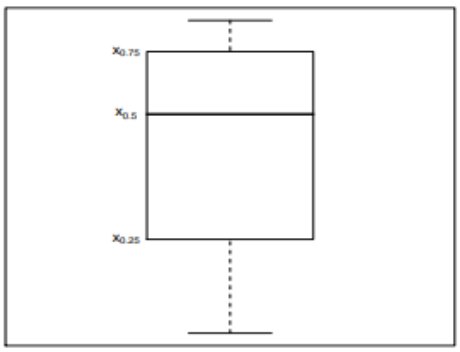
Differenza interquartile

Tale indice pone in evidenza il 50% dei valori più interni della nostra distribuzione maggiormente addensati intorno alla mediana. Poiché tra il primo ed il terzo quartile è sempre incluso il 50% dei dati, più la differenza interquartile (e/o lo scarto interquartile) è piccola, più significa che questo 50% dei dati si concentra attorno alla mediana.

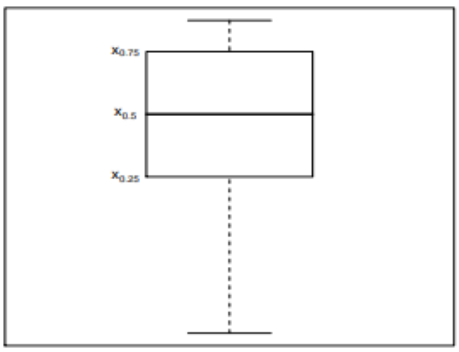
IQR = X0,75 − X0,25

I quartili possono essere rappresentati tramite un grafico detto boxplot. Di seguito riportiamo due boxplot in cui la mediana coincide, ma la distanza tra il primo e il terzo quartile è diversa. Nel grafico di destra la differenza interquartile è più bassa, denotando una maggiore concentrazione del 50% dei dati centrali attorno alla mediana. La mediana è dunque un indicatore sintetico più preciso nel caso B che nel caso A.

A



B



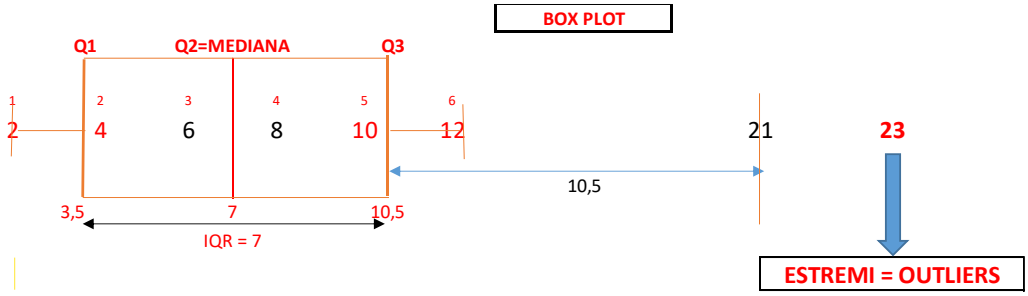
Nota bene: I quantili (e quindi anche i quartili) non possono essere calcolati per i caratteri qualitativi sconnessi: dalla definizione di quantile si intuisce la necessità che il carattere studiato sia misurato almeno su una scala ordinale, ossia abbia modalità ordinabili.

Caratteristiche del boxplot

- Il lato inferiore della box corrisponde al 1° quartile
- Il lato superiore della box corrisponde al 3° quartile
- Il boxplot è costruito rispetto alla mediana
- Il baffo superiore viene tagliato sul valore più alto della distribuzione (non sempre!)
- Il baffo inferiore viene tagliato sul valore più basso della distribuzione (non sempre!)

Nota bene: la lunghezza dei baffi della boxpolt non possono essere più lunghi di 1 volta e mezzo la misura della box corrispondente alla differenza interquartile (Es. se IQR = 12, i baffi potranno essere lunghi al massimo 18 ossia IQR = 12 + IQR/2 = 6.

Nota bene: i valori che eventualmente dovessero risultare esterni ai baffi, sarebbero da considerare valori estremi detti anche valori outliers (quei valori che rendono asimmetrica la distribuzione)



POS Q3 = [(6+1)3]/4 = 5,25
POS Q1 = [(6+1)1]/4 = 1,75
Q3 = 10,5
Q1 = 3,5
IQR = 10,5 - 3,5 = 7

Scostamento semplice medio – Varianza – Deviazione Standard

Tali indici di variabilità pongono in evidenza la dispersione dei dati della distribuzione rispetto alla media.

X	fi	Fi	X fi
6	10	10	60
8	2	12	16
10	4	16	40
12	4	**20**	48
			164

RANGE = 12 - 6 = 6
MEDIA = 8,2
MODA = 6
MEDIANA = 6
SSM = 2,24
S^2 = 5,96
S = 2,44

DEVx = ∑(X-Xmedia)²fi (devianza di x)

media = [(6x10)+(8x2)+(10x4)+12x4)]/20
dev = ∑(Xi-Xmedia)²fi
dev = (6-8,2)²10+(8-8,2)²2+(10-8,2)²4+(12-8,2)²4=119,2
s^2 = dev/n = 119,2/20 = 5,96
s = √varianza = √5,96 = 2,44

$S^2 = \sigma^2$	varianza
$S = \sigma$	dev.standard
$\bar{X} = \mu$	media

CAMPIONE — POPOLAZIONE

SSM = ∑|Xi-Xmedia|fi / n → **S^2 = ∑(Xi-Xmedia)²fi / n** → **S = √∑(Xi-Xmedia)²fi / n**

SSM = $\frac{|6-8,2|10 + |8-8,2|2 + |10-8,2|4 + |12-8,2|4}{20}$ → $\frac{22 + 0,4 + 7,2 + 15,2}{20} = \frac{44,8}{20}$ → 2,24

DEV = ∑(Xi-Xmedia)²fi

S^2 = $\frac{(6-8,2)^2 10 + (8-8,2)^2 2 + (10-8,2)^2 4 + (12-8,2)^2 4}{20}$ → $\frac{48,4 + 0,08 + 12,96 + 57,76}{20} = 5,96$ → **S = √5,96 = 2,44**

la dev.st indica la dispersione dei dati nella distribuzione

Coefficiente di variabilità

Il coefficiente di variazione di un fenomeno X, indicato con CV (X), è un indice di variabilità relativa dato dal rapporto tra lo scarto quadratico medio ed il valore assoluto della media aritmetica, ossia:

$$CV(X) = \frac{\tilde{s}}{|\bar{x}|}.$$

Poiché deviazione standard e media sono espressi nella stessa unità di misura, il coefficiente di variazione non dipende più dall'unità di misura usata per il fenomeno studiato, ossia è un numero puro. Per questa ragione il coefficiente di variazione si usa per confrontare la variabilità dello stesso fenomeno in popolazioni diverse o la variabilità di fenomeni diversi.

I DISTR. **8,05 ± 2,61**

II DISTR. **23,62 ± 3,67**

μ ± σ

CV = (σ/μ)100 → $\frac{2,61}{8,05} \times 100 = 32,42\%$

CV = (σ/μ)100 → (3,67/23,62)100 = 15,53%

Esempio di esercizio

X	fi	Fi	Xc	Xcfi
80--89	8	8	84,5	676
90--99	7	15	94,5	661,5
100--109	1	16	104,5	104,5
110--120	4	**20**	115	460
				1902

Xmedia = 1902/20 = 95,1

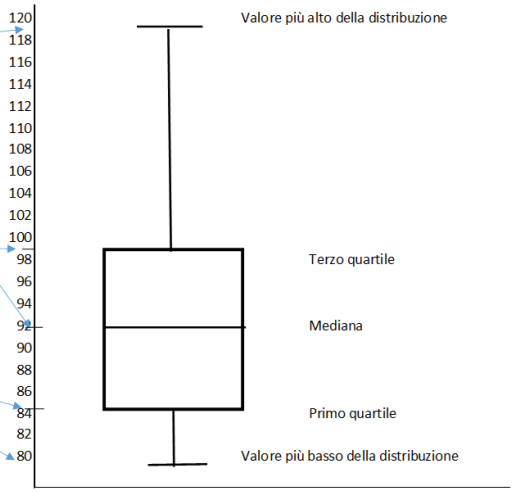
Xmediana = Pos.mediana = 10 classe mediana è la (90--99) ⟶ $89,5 + \frac{(10 - 8)\ 10}{7} = 92,35$

Pos.Q3 = 15 ⟶ Q3 = 99,5

Pos.Q1 = 5 ⟶ Q1 = 85,75

IQR = 99,5 - 85,75 = 13,75

$S = \sqrt{\frac{898,88 + 2,52 + 88,36 + 1584,04}{20}} = 11,34$



Capitolo V – Dai punti Z agli indici di forma

L'analisi e l'elaborazione dei risultati di un test o di una ricerca, al fine di poter essere correttamente interpretati e confrontati con outcome di test/ricerche effettuate su diverse distribuzioni, hanno necessità di essere standardizzati ossia trasformati in punteggi standard.

In tal senso possiamo affermare che il punteggio/risultato di un test/ricerca sarà identificato quale "punteggio grezzo"; l'attributo grezzo sta ad indicare che con tali punteggi occorre compiere ulteriori elaborazioni (procedure di standardizzazione) che permettono confronti con esiti conseguiti in altre prove dagli stessi soggetti o da soggetti diversi.

La trasformazione dei punteggi rende quindi paragonabili dati appartenenti a distribuzioni diverse.

Nel nostro corso indicheremo due metodi di trasformazione:

- I punti Z
- I punti T (trasformazione lineare dei punti Z… T = 50 + 10Z)

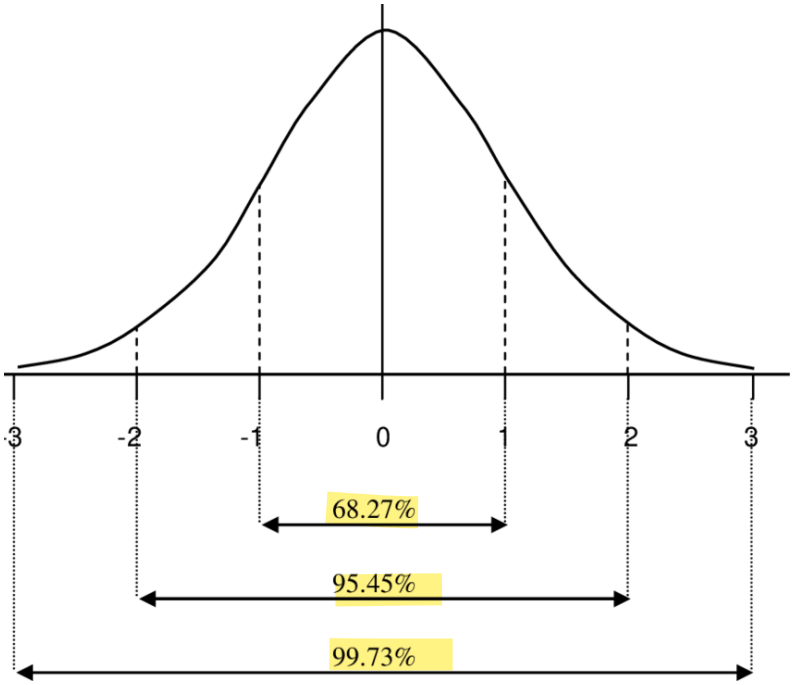
Il punto standard, detto anche punto Z, punteggio standardizzato o punteggio tipificato, definisce la posizione di un soggetto all'interno della sua distribuzione in termini di "quanti scarti tipo sopra o sotto la media" il soggetto si trova. In sostanza si confronta il dato in termini di performance. Viene calcolato con la formula:

$$z = \frac{X - \bar{X}}{s}$$

Se Z è positivo il dato "grezzo" (non standardizzato) del soggetto è superiore alla media, viceversa per Z negativo. Se Z è vicino allo zero, il dato "grezzo" del soggetto è vicino alla media della distribuzione. La media della distribuzione di tutti i punti Z è pari a zero e lo scarto quadratico medio è pari a 1.

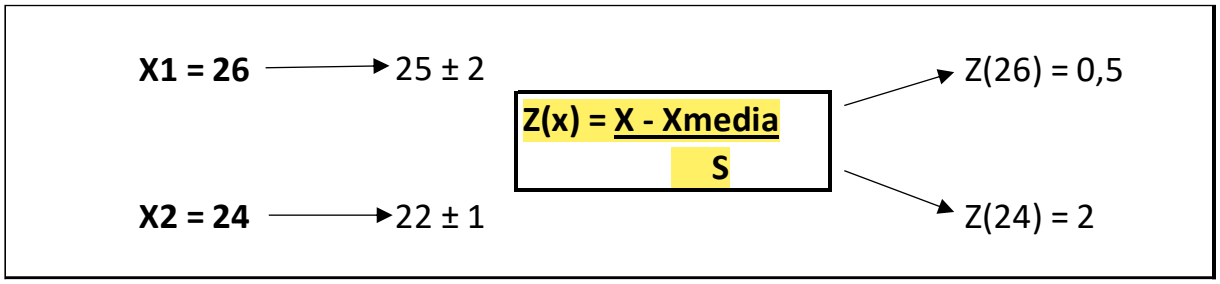
I punteggi Z, a differenza dei valori originari, permettono di confrontare la posizione di un qualunque soggetto nella distribuzione di un carattere, con la posizione dello stesso soggetto nella distribuzione di un altro carattere o la posizione di diversi soggetti in differenti distribuzioni.

Nel caso in cui si trasformano i punteggi osservati in punteggi Z si ottiene una particolare curva che ha media uguale a 0 ($\mu = 0$) e deviazione standard uguale a 1 ($\sigma = 1$) ed è chiamata distribuzione normale standardizzata.



Esempio

All'interno di un corso di studi, abbiamo due studenti (Stefano = X1 e Daniela = X2) che, nel sostenere l'esame di statistica su due diversi appelli, hanno conseguito quale votazione, rispettivamente 26 e 24. Chi tra Stefano e Daniela ha avuto una migliore performance rispetto agli studenti che hanno sostenuto l'esame con loro?



Guardando il valore grezzo, potremmo dire che, in termini assoluti, Stefano ha avuto una prestazione migliore rispetto a Daniela; ma tali dati, appartenendo a distribuzioni differenti, risultano non direttamente paragonabili in quanto caratterizzati da medie e deviazioni standard diverse.

Per poter rispondere alla nostra domanda, abbiamo standardizzato i due valori grezzi (X = 26 e X = 24) per poter poi confrontare i rispetti valori Z.

Come possiamo leggere, Z(x1) = 0,5 e Z(x2) = 2. Possiamo interpretare tale risultato affermando che Stefano ha avuto una performance di 0,5 deviazioni standard migliore rispetto la media dei voti conseguiti nel suo appello mentre Daniela ha avuto una performance di ben 2 deviazioni standard migliore rispetto la media dei voti conseguiti nel

suo appello. In maniera intuitiva possiamo dire che Daniela ha avuto una migliore performance rispetto a Stefano.

Ma come potremmo interpretare in termini quantitativi questi punti Z? Cosa vuol dire che Stefano ha avuto una performance di 0,5 deviazioni standard migliore rispetto la media dei voti conseguiti nel suo appello. Per significare quanto sopra dovremo approfondire la così detta "Distribuzione Normale" o "Gaussiana" e la "Normale Standardizzata".

I punteggi Z vantaggi

- I punti Z al di sotto della media avranno segno negativo, quelli corrispondenti alla media saranno pari a 0 e quelli superiori alla media avranno segno positivo
- L'unità di misura utilizzata è pari ad una deviazione standard
- I punti Z altrimenti detti standard sono posizionati su una scala ad intervalli: è possibile quindi eseguire operazioni matematiche su di essi

I punteggi Z svantaggi

- Non è possibile affidarsi completamente ai punti Z su comparazioni tra prove diverse se la popolazione di riferimento degli studenti considerati non presenta uguali medie e dispersioni negli obiettivi cognitivi considerati
- Il segno negativo è una complicazione nei calcoli
- La deviazione standard è una unità di misura piuttosto ampia, se si considera che in una distribuzione normale il 68 % dei punteggi (ossia dal 16 all'84 centile) è compreso tra -1σ e + 1σ dalla media

I punti T

I punti T assegnano convenzionalmente valore 50 alla media e alla deviazione standard valore 10.

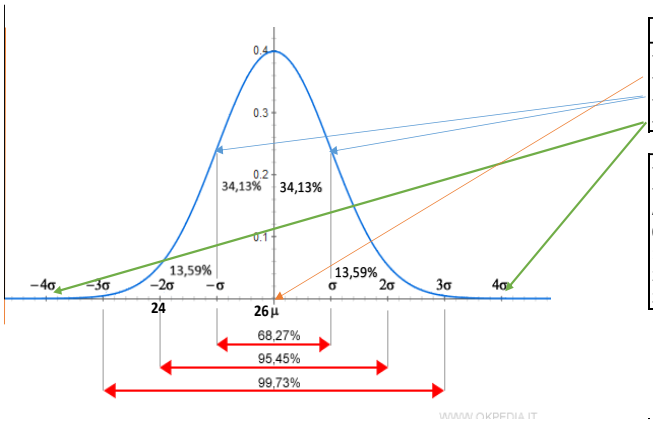
La denominazione "punti T" venne data da W.A. McCall nel 1922 in onore del suo professore E.L. Thorndike (cfr. W.A. McCall, How to measure in Education NY, Macmilian, 1922).

$$T = 50 + 10\frac{x - \bar{x}}{s}$$

La Curva Normale o Gaussiana (Caratteristiche e regola empirica)

La curva gaussiana, nota anche come distribuzione normale, prende il nome dal matematico tedesco Carl Friedrich Gauss ed è diventata un pilastro essenziale dell'analisi statistica venendo utilizzata in diverse discipline.

La curva gaussiana, o distribuzione normale, è una funzione matematica che descrive il modo in cui i dati sono distribuiti in un insieme. È caratterizzata da una forma simmetrica e dal famoso aspetto a campana, con la maggior parte dei dati concentrati intorno alla media, i quali si disperdono uniformemente verso i lati. La funzione è definita da due parametri principali: la media, che rappresenta il centro della distribuzione e la deviazione standard, che misura la dispersione dei dati.



CURVA NORMALE O GAUSSIANA

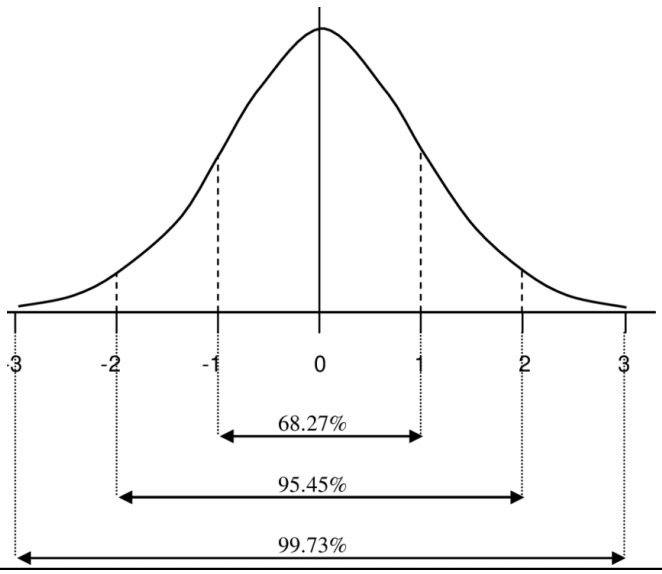
- FORMA CAMPANULARE
- MODA, MEDIA E MEDIANA COINCIDONO DIVIDENDO L'AREA DELLA CURVA IN 2 PARTI UGUALI E SIMMETRICHE
- PUNTO DI FLESSO CADE SULLA I DEVIAZIONE STANDARD
- CODE SONO ASINTOTICHE

- LA CURVA NORMALE STANDARDIZZATA è UNA CURVA NORMALE AVENTE MEDIA 0 E DEVIAZIONE STAND. 1
- REGOLA EMPIRICA: IN UNA DISTRIBUZIONE NORMALE L'AREA COMPRESA ALL'INTERNO DELLA CURVA ASSUME PROPORZIONI DEFINITE E FISSE:

0 < 34,13% < 1
1 < 13,59% < 2
2 < 2,14% < 3
3 < 0,13% < 4

La Curva Normale standardizzata

La curva normale standardizzata, o distribuzione Z, è una distribuzione normale con media uguale a 0 e deviazione standard pari a 1. Questa standardizzazione permette di confrontare facilmente dati provenienti da diverse distribuzioni normali, rendendoli confrontabili perché la forma della curva rimane invariata, cambia solo l'unità di misura.



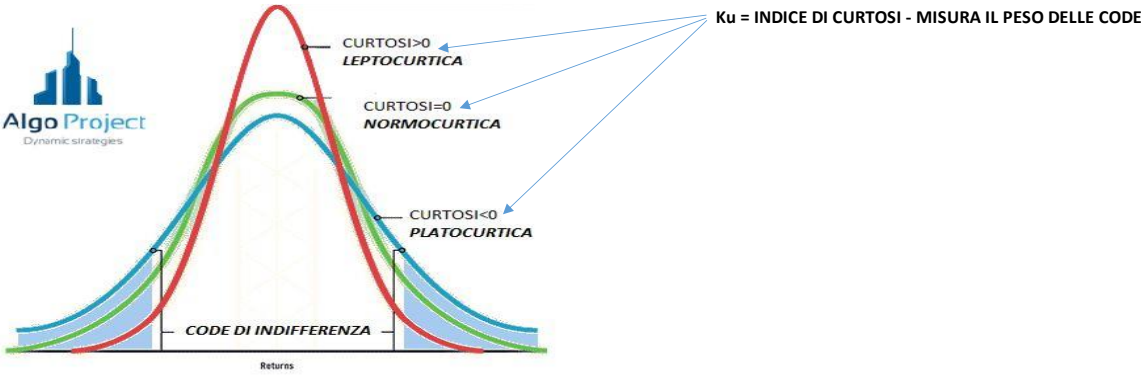
La curva Normale Standardizzata ha le medesime caratteristiche della Gaussiana, avendo però media = 0 e deviazione standard = 1

Al fine di interpretare e correttamente significare i punteggi Z, faremo riferimento alla Tavola della Normale standardizzata.

Indici di Forma

- Indice di curtosi: pone in evidenza il peso delle code.

Nota: più la distribuzione si appiattisce (platocurtica), più aumenta la dispersione dei dati rispetto alla media.



- Indice di asimmetria

In teoria delle probabilità una distribuzione di probabilità è simmetrica quando la sua funzione di probabilità *P* (nel caso discreto) o la sua funzione di densità di probabilità (nel caso continuo) siano simmetriche rispetto ad un particolare valore.

Esempi di distribuzioni simmetriche sono le distribuzioni uniformi (discreta e distribuzione continua uniforme) su insiemi simmetrici, la distribuzione normale e altre distribuzioni

derivate da distribuzioni simmetriche (la distribuzione t di Student) oppure definite in maniera simmetrica (la distribuzione di Skellam con parametri uguali).

Un indice di asimmetria (in inglese *skewness*) di una distribuzione è un valore che cerca di fornire una misura della sua mancanza di simmetria.

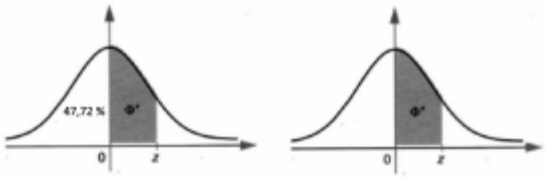
Nota: in una distribuzione asimmetrica a destra, la media assume valori maggiori rispetto alla mediana, contrariamente alla distribuzione asimmetrica a sinistra, nella quale la media assumerà valori minori rispetto alla mediana.

Tavola della normale standardizzata

Distribuzione normale $\phi^*(z)$

Area sottesa alla curva normale standardizzata da 0 a z
- test a due code: $z_{\alpha} = 1{,}96$; $z_{\alpha} = 2{,}58$
- test a una coda: $z_{\alpha} = 1{,}65$; $z_{\alpha} = 2{,}33$

La cifra in testa alla colonna rappresenta la seconda cifra decimale di z

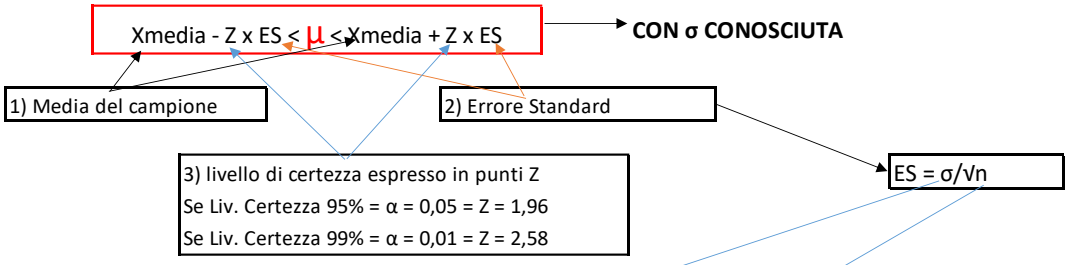


z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0754
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1809	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4811
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	,0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999

Capitolo VI – Accenni all'intervallo di fiducia o confidenza

In statistica, quando si stima un parametro, è spesso insufficiente individuare un singolo valore: è opportuno allora accompagnare la stima con un intervallo di valori probabili per quel parametro, definito **intervallo di confidenza** (o **intervallo di fiducia**, o **intervallo fiduciario**).

A questo intervallo di confidenza si associa quindi un valore di probabilità cumulativa che caratterizza, indirettamente in termini di probabilità, la sua ampiezza rispetto ai valori massimi assumibili dalla variabile aleatoria. Cioè il valore di probabilità cumulativa indica la probabilità che l'evento casuale descritto dalla variabile aleatoria cada all'interno di suddetto intervallo di confidenza, graficamente pari all'area sottesa dalla curva di distribuzione di probabilità della variabile aleatoria nell'intervallo considerato.



ESERCIZIO

Avendo una popolazione di 100 persone, stimarne il peso medio, avendo un campione composto da **10 soggetti** del peso di:

Sog.	1	2	3	4	5	6	7	8	9	10
Peso Kg	70	75	76	77	78	80	82	84	86	68

e sapendo che la **σ = 2** ed **α = 0,01**

Xmedia = 77,6
ES = 0,63
Z = 2,58

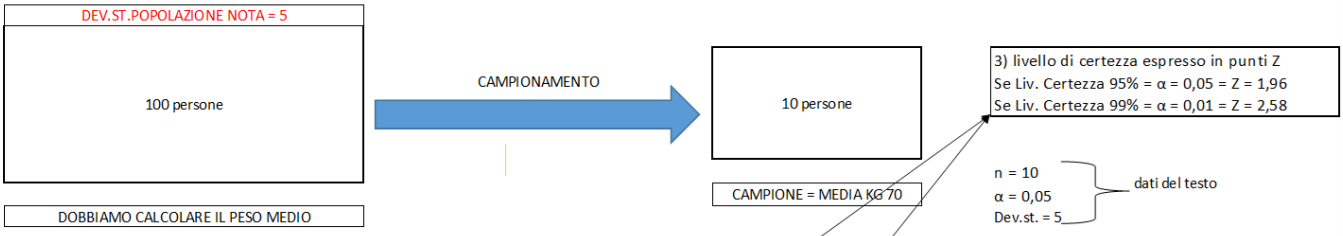
77,6 - 2,58 x 0,63 < μ < 77,6 + 2,58 x 0,63 → **75,98 < μ < 79,22**

La media della pop. È compresa tra 75,98 e 79,22 con un liv. Di certezza del 99%

Xmedia - T x ES stimato < μ < Xmedia + T x ES stimato → CON σ SCONOSCIUTA CON N<30

ES Stimato = S/√n

Xmedia - Z x ES stimato < μ < Xmedia + Z x ES stimato → CON σ SCONOSCIUTA CON N>30



1) PESIAMO TUTTI I SOGGETTI PER CALCOLARE LA MEDIA ... NON POSSIAMO FARLO X ASSENZA DI RISORSE
2) ESTRAIAMO UN CAMPIONE RAPPRESENTATIVO E PROBABILISTICO DEL QUALE CALCOLARE LA MEDIA DEL PESO
3) CALCOLO LA MEDIA DEL CAMPIONE = KG 70
4) POSSIAMO AFFERMARE CHE LA MEDIA DELLA POPOLAZIONE SIA DI KG 70 PRECISI? NO... PERCHE' ESISTE UNA CERTA DIFFERENZA NUMERICA!
5) PARTENDO DALLA MEDIA DEL CAMPIONE, POSSIAMO **STIMARE** LA MEDIA DELLA POPOLAZIONE UTILIZZANDO L'INTERVALLO DI CONFIDENZA O FIDUCIA
6) ESISTONO 3 TIPI DI INTERVALLO DI FIDUCIA: 1° CON DEV.ST. DELLA POPOLAZIONE NOTA; 2° CON DEV.ST.POPOL. SCONOSCIUTA ED N<30; 3° CON DEV.ST.POP. SCONOSCIUTA ED N>30
7) CONOSCENDO LA DEV.ST.DELLA POPOLAZIONE CALCOLIAMO IL 1°: Xmedia - Z x ES < μ < Xmedia + Z x ES
8) PER CALCOLARE L'INTERVALLO ABBIAMO BISOGNO DI 3 ELEMENTI: MEDIA DEL CAMPIONE, PUNTO Z DATO DAL LIVELLO DI CERTEZZA, ERRORE STANDARD
9) CALCOLO L'INTERVALLO: **70 - 1,96x(5/√10) < μ < 70 + 1,96(5/√10)** 66,9 < μ < 73,1

Xmedia - T x ES stimato < μ < Xmedia + T x ES stimato → CON σ SCONOSCIUTA CON N<30

ES Stimato = S/√n

Xmedia - Z x ES stimato < μ < Xmedia + Z x ES stimato → CON σ SCONOSCIUTA CON N>30

Capitolo VII – La statistica bivariata

Frequentemente due o più caratteri vengono rilevati congiuntamente (contemporaneamente) su “n” unità statistiche. In questo contesto l'obiettivo principale della statistica è indagare gli eventuali legami tra i caratteri rilevati.

In questo corso si considera solo il caso di due variabili (quantitative o categoriali) rilevate sulle stesse unità statistiche, ma gli strumenti statistici descritti possono essere opportunamente estesi allo studio di più di due variabili. Quando si studiano congiuntamente due caratteri, il primo obiettivo è stabilire se esiste un qualche legame tra i due. A questo scopo in letteratura sono stati proposti svariati indici:

- L'indice Chi quadrato di Pearson normalizzato;
- Dipendenza in media ed Eta quadro;
- Coefficiente di correlazione lineare di Pearson atto a misurare la correlazione lineare tra due variabili quantitative.

Ovviamente prima è necessario operare una sintesi dei dati riclassificandoli in forma tabellare.

Tabelle di contingenza: frequenze assolute e relative

Siano X e Y due caratteri rilevati congiuntamente su “n” unità statistiche e sia

{(x1, y1),(x2, y2), . . . ,(xi , yi), . . .(xn, yn)}

la serie di dati grezzi ottenuti con questa rilevazione. Supponiamo che i caratteri X e Y si manifestino rispettivamente attraverso h e k modalità distinte: {x1, x2,...,xh} e {y1, y2,...,hk} e che alcune coppie di valori rilevati si ripetano, allora una prima sintesi delle informazioni contenute nei dati è la tabella di contingenza o tabella a doppia entrata: La quantità nij è detta frequenza assoluta congiunta (Fo) ed è il numero di unità statistiche che presentano la coppia di modalità (xi , yj), per i=1,...,h e j=1,...,k.

Le frequenze riportate nell'ultima colonna

$$\boxed{n_{i.} = \sum_{j=1}^{k} n_{ij}},$$

con i=1,...,h, sono dette frequenze marginali di X; ni. rappresenta il numero di unità statistiche che presentano modalità xi del carattere X.

Le frequenze riportate nell'ultima riga

$$\boxed{n_{.j} = \sum_{i=1}^{h} n_{ij}}$$

con j=1,...,k, sono dette frequenze marginali di Y ; n.j rappresenta il numero di unità statistiche che presentano modalità yj del carattere Y.

X	Y					
	y_1	$\dots$	y_j	$\dots$	y_k	$n_{i.}$
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1k}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ik}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_h	n_{h1}	$\dots$	n_{hj}	$\dots$	n_{hk}	$n_{h.}$
$n_{.j}$	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.k}$	n

Tabella 1: Tabella di contingenza $h \times k$

Ovviamente la somma delle frequenze assolute congiunte è pari alla somma delle frequenze assolute marginali ed al totale delle unità statistiche n.

Esempio

Supponiamo che per uno studio sul mercato del lavoro si siano intervistati 20 lavoratori e si siano rilevati congiuntamente i caratteri “classe stipendiale” e “grado di istruzione”, ottenendo la seguente serie di dati:

{(Bassa; Licenza media); (Bassa; Licenza media);(Bassa; Licenza media);(Media; Licenza media);(Alta; Licenza media);(Bassa; Licenza superiore);(Bassa; Licenza superiore);(Media; Licenza superiore); (Media; Licenza superiore);(Media; Licenza superiore);(Media; Licenza superiore);(Alta; Licenza superiore);(Alta; Licenza superiore);(Alta; Licenza superiore);(Bassa; Laurea);(Media; Laurea);(Media; Laurea);(Alta; Laurea);(Alta; Laurea);(Alta; Laurea)}

Il numero di modalità distinte del carattere X = ”classe stipendiale” è h=3;

il numero di modalità distinte del carattere Y = ”grado di istruzione” è k=3;

la corrispondente tabella di contingenza 3 × 3 è

Classe stipendiale (X)	Grado di istruzione (Y)			
	Licenza media	Licenza superiore	Laurea	$n_{i.}$
Bassa	$n_{11}=3$	$n_{12}=2$	$n_{13}=1$	$n_{1.}=3+2+1=6$
Media	$n_{21}=1$	$n_{22}=4$	$n_{23}=2$	$n_{2.}=1+4+2=7$
Alta	$n_{31}=1$	$n_{32}=3$	$n_{33}=3$	$n_{3.}=1+3+3=7$
$n_{.j}$	$n_{.1}=3+1+1=5$	$n_{.2}=2+4+3=9$	$n_{.3}=1+2+3=6$	20

Come per l'analisi di un fenomeno univariato anche nello studio congiunto di due fenomeni possiamo ragionare in termini di proporzioni e fare riferimento alle frequenze relative, date dal rapporto tra le frequenze assolute e il totale delle unità statistiche.

La quantità $\boxed{f_{ij} = n_{ij}/n}$ è detta frequenza relativa congiunta ed indica la proporzione di unità statistiche che presentano la coppia di modalità (xi , yj), per i=1,...,h e j=1,...,k.

Le frequenze $f_{i.} = \sum_{j=1}^{k} f_{ij} = n_{i.}/n$, con i=1,...,h, sono dette frequenze relative marginali di X e rappresentano la proporzione di unità statistiche che hanno modalità xi del carattere X.

Le frequenze $f_{.j} = \sum_{i=1}^{h} f_{ij} = n_{.j}/n$, con j=1,...,k, sono dette frequenze relative marginali di Y e rappresentano la proporzione di unità statistiche che presentano modalità yj del carattere Y.

Ovviamente la somma delle frequenze relative congiunte coincide con la somma delle frequenze relative marginali ed è pari a 1, ossia:

$$\sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} = \sum_{i=1}^{h} f_{i.} = \sum_{j=1}^{k} f_{.j} = 1$$

Esempio

Con riferimento all'esempio precedente, la corrispondente tabella delle frequenze relative è:

Classe stipendiale (X)	Grado di istruzione (Y)			
	Licenza media	Licenza superiore	Laurea	$f_{i.} = n_{i.}/n$
Bassa	$f_{11} = 3/20 = 0.15$	$f_{12} = 2/20 = 0.1$	$f_{13} = 1/20 = 0.05$	$f_{1.} = 6/20 = 0.3$
Media	$f_{21} = 1/20 = 0.05$	$f_{22} = 4/20 = 0.2$	$f_{23} = 2/20 = 0.1$	$f_{2.} = 7/20 = 0.35$
Alta	$f_{31} = 1/20 = 0.05$	$f_{32} = 3/20 = 0.15$	$f_{33} = 3/20 = 0.15$	$f_{3.} = 7 = 0.35$
$f_{.j} = n_{.j}/n$	$f_{.1} = 5/20 = 0.25$	$f_{.2} = 9/20 = 0.45$	$f_{.3} = 6/20 = 0.3$	1

Le frequenze relative condizionate

La prima colonna della tabella a doppia entrata dei dati riporta le modalità distinte del carattere X. Se volessimo studiare il fenomeno X a prescindere dal carattere Y allora faremmo riferimento alla sola distribuzione marginale di X ottenuta dalla prima e dall'ultima colonna della tabella a doppia entrata. Diversamente si potrebbe voler studiare il comportamento del fenomeno X anziché su tutte le unità statistiche, sulle sole nj unità che presentano modalità yj del carattere Y. In questo caso bisognerebbe far riferimento alla tabella di frequenze, detta distribuzione delle frequenze (assolute) di X condizionata a yj:

	Y = ETA'			
X = TIT.STUDIO	**40**	**45**	**50**	**Ri**
LIC. ELEM	4	3	2	9
LIC.MEDIA	6	7	2	15
DIPLOMA	2	6	5	13
LAUR. I LIV	8	4	6	18
LAUREA II	2	5	3	10
MASTER	6	1	2	9
Ci	28	26	20	**74**

X	Frequenze condizionate a y_j
x_1	n_{1j}
$\vdots$	$\vdots$
x_i	n_{ij}
$\vdots$	$\vdots$
x_h	n_{hj}
	$n_{.j}$

Poiché Y si manifesta attraverso k modalità distinte si possono individuare k sottoinsiemi di unità statistiche, uno per ciascuna modalità di Y. La prima colonna della Tabella, unitamente a ciascuna delle colonne centrali, costituiscono le k distribuzioni condizionate di X. Poiché le k sottopopolazioni individuate dalle varie modalità di Y hanno numerosità diversa, rispettivamente n.1, . . . , n.k, per operare confronti (relativi al comportamento di X nei k gruppi) è necessario ragionare in termini relativi; per questa ragione si definiscono le distribuzioni condizionate relative di X come rapporto tra le frequenze assolute e il totale delle unità statistiche nel gruppo di appartenenza. Le colonne centrali della seguente tabella sono le frequenze relative di X, condizionate alle varie modalità di Y (ossia le frequenze relative di X all'interno delle k sottopopolazioni individuate dalle diverse modalità di Y); l'ultima colonna riporta le frequenze relative marginali di X e descrive il comportamento di X in tutte le unità statistiche indagate (sull'intera popolazione di n unità)

Modalità del fenomeno d'interesse X	Freq. relative condizionate a					Freq. relative marginali
	y_1	$\dots$	y_j	$\dots$	y_k	
x_1	$\frac{n_{11}}{n_{.1}}$	$\dots$	$\frac{n_{1j}}{n_{.j}}$	$\dots$	$\frac{n_{1k}}{n_{.k}}$	$f_{1.} = \frac{n_{1.}}{n}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	$\frac{n_{i1}}{n_{.1}}$	$\dots$	$\frac{n_{ij}}{n_{.j}}$	$\dots$	$\frac{n_{ik}}{n_{.k}}$	$f_{i.} = \frac{n_{i.}}{n}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_h	$\frac{n_{h1}}{n_{.1}}$	$\dots$	$\frac{n_{hj}}{n_{.j}}$	$\dots$	$\frac{n_{hk}}{n_{.k}}$	$f_{h.} = \frac{n_{h.}}{n}$
Totali	1	$\dots$	1	$\dots$	1	1

Invertendo i ruoli tra X e Y, si può essere interessati a studiare il fenomeno Y marginalmente ossia su tutte le unità statistiche (a prescindere da X).

In una tabella di contingenza si possono dunque individuare più distribuzioni di frequenza:

- La distribuzione delle frequenze congiunte (Fo);
- Due distribuzioni marginali, una per X e l'altra per Y (Ri e Ci);
- h distribuzioni condizionate di Y da xi con i=1,...,h;
- k distribuzioni condizionate di X da yj con j=1,...,k.

Esempio

Ad un gruppo di 18 fumatori è stato chiesto quanti pacchetti di sigarette hanno fumato nell'ultimo mese (fenomeno X) e quante volte hanno praticato un'attività sportiva (fenomeno Y). I dati sono riportati di seguito:

1. Si costruisca una tabella a doppia entrata riclassificando il fenomeno X nelle seguenti tre classi: $(0;15]$, $(15;30]$ e $(30;45]$ e il fenomeno Y nei seguenti tre intervalli: $(0;7]$, $(7;14]$ e $(14;21]$:

	Y			
X	$(0;7]$	$(7;14]$	$(14;21]$	$n_{i.}$
$(0;15]$	6	3	0	9
$(15;30]$	5	1	1	7
$(30;45]$	1	1	0	2
$n_{.j}$	12	5	1	18

2. Si riportino le distribuzioni marginali (sia assolute che relative) di X e di Y.

   Distribuzioni marginali di X:

X	$n_{i.}$	$f_{i.} = n_{i.}/n$
$(0;15]$	9	$9/18 = 0.50$
$(15;30]$	7	$7/18 = 0.3\bar{8}$
$(30;45]$	2	$2/18 = 0.1\bar{1}$
Totali	18	1

   Distribuzioni marginali di Y:

Y	$n_{.j}$	$f_{.j} = n_{.j}/n$
$(0;7]$	12	$12/18 = 0.6\bar{6}$
$(7;14]$	5	$5/18 = 0.2\bar{7}$
$(14;21]$	1	$1/18 = 0.0\bar{5}$
Totali	18	1

3. Si riportino tutte le distribuzioni condizionate (assolute e relative) di X e Y.

   Distribuzioni condizionate (assolute) di X (si ottengono leggendo per colonne la tabella a doppia entrata dei dati):

X	Modalità di Y (gruppi) $(0;7]$	$(7;14]$	$(14;21]$
$(0;15]$	6	3	0
$(15;30]$	5	1	1
$(30;45]$	1	1	0
Dimensione gruppi	12	5	1

Distribuzioni condizionate (relative) di X:

X	Modalità di Y (gruppi) $(0;7]$	$(7;14]$	$(14;21]$
$(0;15]$	$6/12 = 0.500$	$3/5 = 0.6$	0
$(15;30]$	$5/12 = 0.16\bar{6}$	$1/5 = 0.2$	1
$(30;45]$	$1/12 = 0.08\bar{3}$	$1/5 = 0.2$	0
Totali	1	1	1

Distribuzioni condizionate (assolute) di Y (si ottengono leggendo per righe la tabella a doppia entrata):

Y	Modalità di X (gruppi) $(0;15]$	$(15;30]$	$(30;45]$
$(0;7]$	6	5	1
$(7;14]$	3	1	1
$(14;21]$	0	1	0
Dimensione gruppi	9	7	2

Distribuzioni condizionate (relative) di Y:

Y	Modalità di X (gruppi) $(0;15]$	$(15;30]$	$(30;45]$
$(0;7]$	$6/9 = 0.6$	$5/7 = 0.714$	$1/2 = 0.5$
$(7;14]$	$3/9 = 0.\bar{3}$	$1/7 = 0.143$	$1/2 = 0.5$
$(14;21]$	0	$1/7 = 0.143$	0
Totali	1	1	1

Indipendenza di due caratteri

Per semplicità espositiva, introduciamo il concetto di indipendenza tra due caratteri attraverso un esempio:

Esempio

Si supponga di aver effettuato un'indagine su 210 lavoratori di età compresa tra i 30 e i 35 anni e di aver rilevato i fenomeni “titolo di studio” (X) e “reddito” (Y). Si supponga che la classificazione dei dati abbia portato alla seguente tabella a doppia entrata:

X	Y (800; 1500]	(1500; 2000]	(2000; 3000]	$n_{i.}$
Media Inferiore	15	12	3	30
Media Superiore	45	36	9	90
Laurea o più	40	32	8	80
$n_{.j}$	100	80	20	200

Potrebbe essere interessante stabilire se il reddito si distribuisce nello stesso modo nei tre diversi livelli d'istruzione. A questo scopo determiniamo le distribuzioni di frequenze condizionate di Y. In particolare, dovendo operare un confronto in tre gruppi di dimensione diversa, facciamo riferimento alle frequenze condizionate relative.

Y	Modalità di X (gruppi) Media Inferiore	Media Superiore	Laurea o più	$f_{i.}$
(800; 1500]	$15/30 = 0.5$	$45/90 = 0.5$	$40/80 = 0.5$	$100/200 = 0.5$
(1500; 2000]	$12/30 = 0.4$	$36/90 = 0.4$	$32/80 = 0.4$	$80/200 = 0.4$
(2000; 3000]	$3/30 = 0.1$	$9/90 = 0.1$	$8/80 = 0.1$	$20/200 = 0.1$
Totali	1	1	1	1

Dalla tabella precedente si evince immediatamente che per ciascun livello d'istruzione la proporzione di lavoratori nelle tre classi di reddito è la stessa. Questo significa che avere un'istruzione diversa non comporta una distinta distribuzione del reddito. In termini statistici si dice che la distribuzione del reddito non è influenzata dal livello d'istruzione e questo concetto si esprime semplicemente affermando che il reddito è indipendente dal titolo di studio. Si noti che in caso di indipendenza tra due caratteri le distribuzioni relative condizionate coincidono con la distribuzione relativa marginale. Allo stesso modo si potrebbe essere interessati allo studio del livello d'istruzione in ciascuna classe di reddito. A tale scopo si determinano le distribuzioni condizionate relative di X esattamente come precedentemente fatto con le distribuzioni condizionate di Y.

Connessione e l'indice Chi quadrato

Si dice che esiste un legame tra due caratteri X e Y se essi non sono indipendenti e la presenza di un legame è detta connessione. La condizione opposta rispetto all'indipendenza è la massima connessione. Due caratteri X e Y si dicono massimamente connessi se data una modalità di uno dei due caratteri, l'altro può assumere una ed una sola modalità. In altri termini, c'è massima connessione tra X e Y se c'è dipendenza funzionale esatta di un carattere dall'altro.

Solitamente una tabella di contingenza costruita su dati reali non è né di massima connessione né di indipendenza. Quando non c'è indipendenza si dice che c'è connessione tra i due caratteri studiati ed è interessante stabilire il grado di connessione esistente. In modo semplicistico si può affermare che il legame tra X e Y è tanto è più forte quanto più la tabella osservata dei dati è lontana dalla tabella teorica di indipendenza e vicina ad una tabella di massima connessione. Quindi per misurare il grado di connessione si usa una sorta di misura di distanza tra tabella osservata e tabella teorica. La differenza nij − nˆij è detta contingenza (i, j)-esima e misura la distanza tra la frequenza osservata e quella attesa in ipotesi di indipendenza. Ovviamente c'è indipendenza solo se tutte le h × k contingenze sono nulle ossia se si osservano le frequenze teoriche. Una misura di distanza complessiva tra la tabella osservata e la tabella teorica è l'indice X2 di Pearson:

$$X^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

ossia

$$\chi^2 = \sum \frac{(Fo - Fe)^2}{Fe}$$

che tiene conto di tutte le contingenze.

Ovviamente il valore minimo del Chi Quadrato è lo zero, che viene assunto solo quando tutte le contingenze sono nulle ossia se c'è indipendenza. Si può dimostrare inoltre che l'indice Chi Quadrato assume il suo massimo valore (max X2) solo in presenza di massima connessione e che max X2=n · min{h − 1; k − 1}.

Come già sottolineato in precedenza, nei casi reali non si osservano quasi mai tabelle di massima connessione o di indipendenza e per stabilire se la tabella osservata è prossima a quella teorica o ad una di massima connessione conviene fare riferimento all'indice di Pearson normalizzato:

$$\tilde{X}^2 = \frac{X^2}{\max X^2} = \frac{\sum_{i=1}^{h} \sum_{j=1}^{k} \frac{n_{ij}}{n_{i\cdot} \cdot n_{\cdot j}} - 1}{\min\{h-1; k-1\}}$$

che rispetto al Chi Quadrato è più facilmente interpretabile poiché 0 ≤ X2 ≤ 1. Lo zero corrisponde all'indipendenza e l'uno alla massima connessione; più X2 è vicino a 0 più X e Y sono vicini alla condizione di indipendenza, viceversa più X2 è vicino a 1 più è forte il legame tra i due caratteri. Si noti che l'idea di vicinanza a zero o ad uno è piuttosto soggettiva.

CALCOLO χ^2	$\chi^2 = \sum \frac{(Fo - Fe)^2}{Fe}$

GDL = (R-1)(C-1)

$Fe = \frac{Ri \times Ci}{n}$

Fo			
Y \ X	M	F	Ri
L.E.	4	8	12
L.M.	6	4	10
DIPL	12	6	18
LAUR	2	8	10
Ci	24	26	50

Fe			
	M	F	Ri
L.E.	5,76	6,24	12
L.M.	4,8	5,2	10
DIPL	8,64	9,36	18
LAUR	4,8	5,2	10
Ci	24	26	50

TABELLA DI INDIPENDENZA

$\chi^2 = (4-5,76)^2/5,76 + (8-6,24)^2/6,24 + (6-4,8)^2/4,8 + (4-5,2)^2/5,2 + (12-8,64)^2/8,64 + (6-9,36)^2/9,36 + (2-4,8)^2/4,8 + (8-5,2)^2/5,2 = 7,28$

Esempio

con riferimento agli stessi dati del precedente esempio, per calcolare l'indice X2 (Chi Quadrato) bisogna determinare tutte le contingenze, che vengono riportate nella seguente tabella:

$n_{ij} - \hat{n}_{ij}$	$(18; 22]$	$(22; 26]$	$(27; 30]$	Somme
Istituto tecnico	10-9.24=0.76	15-15.4=-0.4	17-17.36=-0.36	0
Liceo	-7.1	-0.17	7.23	0
Altro	6.34	0.57	-6.91	0
Somme	0	0	0	0

Si noti che la somma delle contingenze (sia per riga che per colonna) è sempre nulla; ciò è dovuto al fatto che le frequenze marginali sono le stesse sia nella tabella osservata che nella tabella teorica.

Per cui

$$X^2 = 0.0625 + 0.0104 + \cdots + 2.1793 = 12.1910$$

Per determinare l'indice di Pearson normalizzato è necessario calcolare il massimo di X^2:

$$\max X^2 = n \cdot \min\{h-1; k-1\} = 150 \cdot \min\{2; 2\} = 150 \cdot 2 = 300,$$

da cui

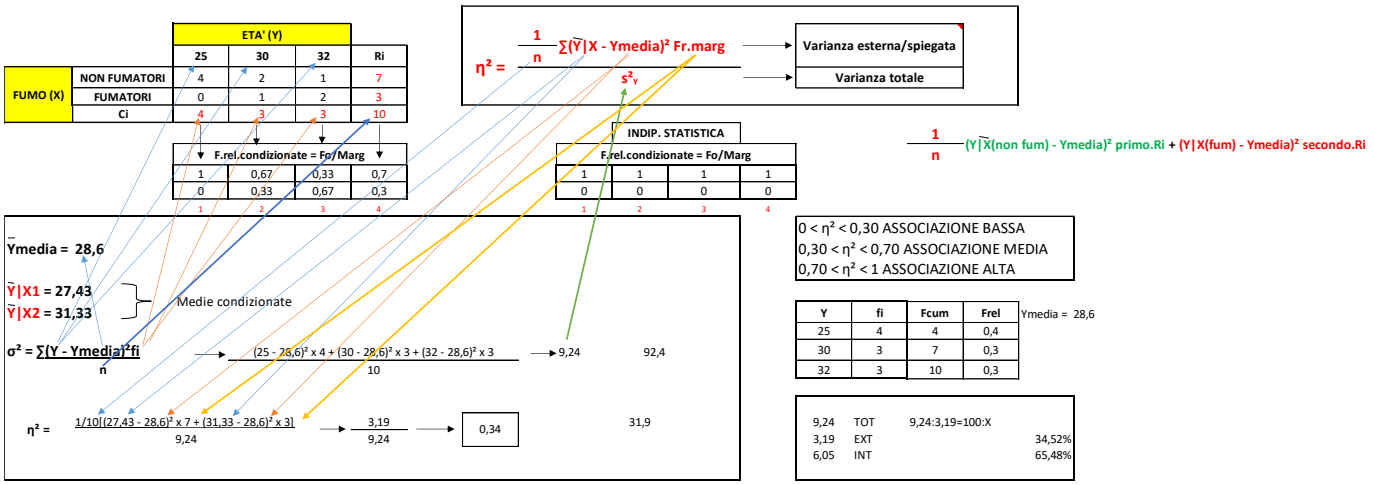
$$\tilde{X}^2 = \frac{X^2}{\max X^2} = \frac{12.1910}{300} = 0.0406.$$

Essendo X2=0.0406 un valore molto prossimo allo zero, possiamo concludere che sulla base dei dati analizzati sembra non esserci alcun legame tra “voto medio” e “istituto di provenienza”.

La dipendenza in media

L'indice X2 di Pearson normalizzato serve a stabilire se esiste un qualche legame tra i caratteri X e Y e a misurarne l'intensità. Qualora si possa concludere che X e Y sono

connessi, si può voler indagare sul tipo di legame che intercorre tra X e Y. Il primo passo è quello di dare un ruolo specifico alle due variabili, che non possono più essere trattate in modo simmetrico. Una deve svolgere il ruolo di variabile esplicativa (che spiega) o indipendente, l'altra quello di variabile spiegata (o dipendente). É la natura del problema oggetto di studio e la logica che suggeriscono il ruolo delle due variabili. Chiamiamo Y la variabile dipendente e X quella esplicativa. Se Y è quantitativa una tipologia di legame che possiamo indagare è la dipendenza in media di Y da X. In altri termini, si vuole stabilire se la media di Y cambia significativamente a seconda delle modalità (numeriche o meno) assunte da X.



		TITOLO DI STUDIO (X)						
		L.MED	DIPL	LAUREA	MASTER	Ri	Yc	Yc Ri
REDDITO (Y)	1--5	41	21	10	24	96	3	288
	5--15	22	13	43	12	90	10	900
	15--30	10	50	61	7	128	22,5	2880
	30--50	9	71	19	11	110	40	4400
	Ci	82	155	133	54	424		8468

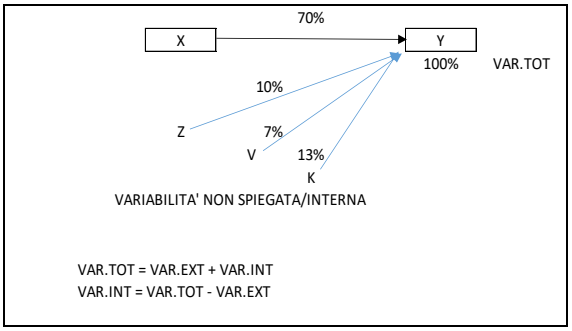
$S^2 = \frac{\sum(Yc - Ymedia)^2 fi}{n}$

Y	Fi	Fcum	Yc	YcFi
1--5	96	96	3	288
5--15	90	186	10	900
15-30	128	314	22,5	2880
30--50	110	**424**	40	4400

Ymedia = 19,97

Ymedia = 19,97
Y|X1 = 11,31
Y|X2 = 26,82
Y|X3 = 19,49
Y|X4 = 14,62
$\sigma^2 y$ = 192,32
η^2 = 0,18
$\sigma^2_{y|x1}$ = 113,652
$\sigma^2_{y|x2}$ = 186,19
$\sigma^2_{y|x3}$ = ...
$\sigma^2_{y|x4}$ = ...

$\frac{1}{424}(11,31 - 19,97)^2 82 + (26,82 - 19,97)^2 155 + (19,49 - 19,97)^2 133 + (14,62 - 19,97)^2 54$



TEOREMA DELLA DEVIANZA

$$\eta^2 = \frac{BSS}{TSS}$$

X: PROGRAMMA DI STUDIO (NOMINALE)

Y: APPRENDIMENTO MISURATO ATTRAVERSO IL VOTO (CARDINALE)

SPE	ORD
27	16
21	24
	23

	VOTO					
PR.ST.	16	21	23	24	27	Ri
SPE	0	1	0	0	1	2
ORD	1	0	1	1	0	3
Ci	1	1	1	1	1	5

TSS = BSS + WSS

TSS: DEVIANZA TOTALE

BSS: DEVIANZA ESTERNA (SPIEGATA)

WSS: DEVIANZA INTERNA (NON SPIEGATA)

$TSS = \sum(Y - Ymedia)^2 fi$

$BSS = \sum(Y|x - Ymedia)^2$(numero di soggetti di ciascuna categoria)

Ymedia =	22,2
Y|x1 =	24
Y|x2 =	21
TSS =	66,8
BSS =	10,8
WSS =	56
η^2 =	0,16

Negli esempi sopra riportati, si rende necessario:

- Calcolare le medie condizionate
- Se queste risultano tra loro uguali avremo una indipendenza in media
- Se queste risultano tra loro diverse avremo una dipendenza in media
- Avendo un condizionamento in media, procediamo con la quantificazione dell'associazione, calcolando eta quadro.

Dipendenza lineare, Coefficiente di Correlazione lineare di Pearson e retta di regressione lineare secondo il metodo dei minimi quadrati

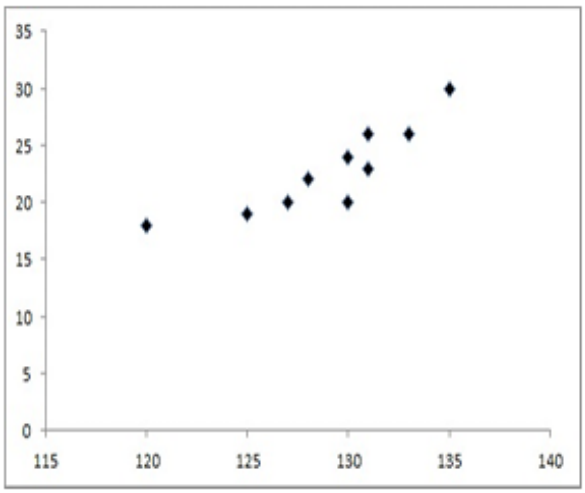
Siano X e Y due caratteri quantitativi; in questo contesto, specialmente se uno dei due caratteri è quantitativo continuo, la serie di dati grezzi (X1; Y1) (X2; Y2) (Xn; Yn) è costituita da coppie di valori distinte, per cui non è possibile una prima sintesi dei dati attraverso una riclassificazione in una tabella di continenza. Lo scopo dell'analisi di correlazione è studiare l'eventuale relazione lineare tra i due fenomeni X e Y.

Il primo strumento utile a questo scopo è il diagramma a dispersione. Un diagramma a dispersione è un grafico in cui si rappresentano i valori che le variabili X e Y assumono sulle n unità statistiche. Una variabile (ad esempio la X) viene rappresentata nell'asse delle ascisse e l'altra (Y) viene rappresentata nell'asse delle ordinate, facendo corrispondere un punto ad ogni coppia di valori osservati (xi yi), per i = 1,…, n.

Per esempio supponiamo di aver rilevato l'altezza (X) e il peso (Y) di 10 bambini di 10 anni, ottenendo i seguenti valori:

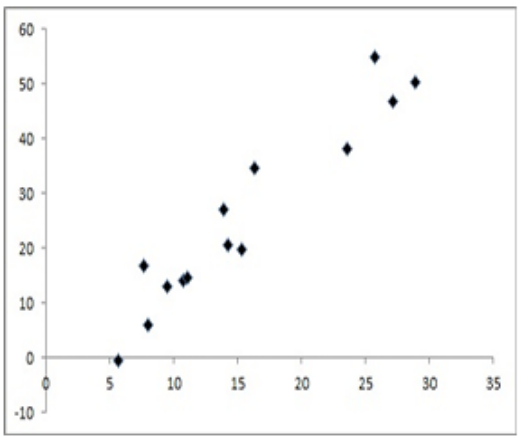
x_i	y_i
120	18
125	19
127	20
128	22
130	20
130	24
131	23
131	26
133	26
135	30

Il diagramma a dispersione corrispondente è:

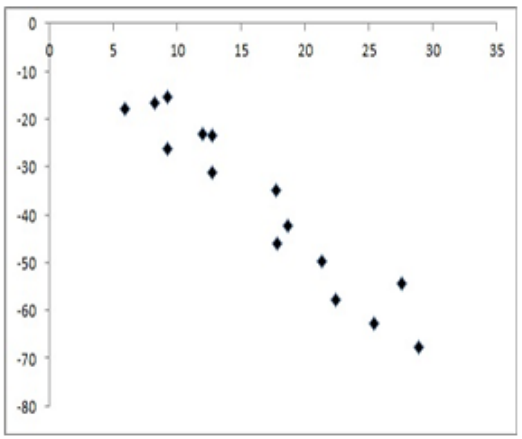


I seguenti grafici a dispersione evidenziano una relazione lineare tra X e Y:

A



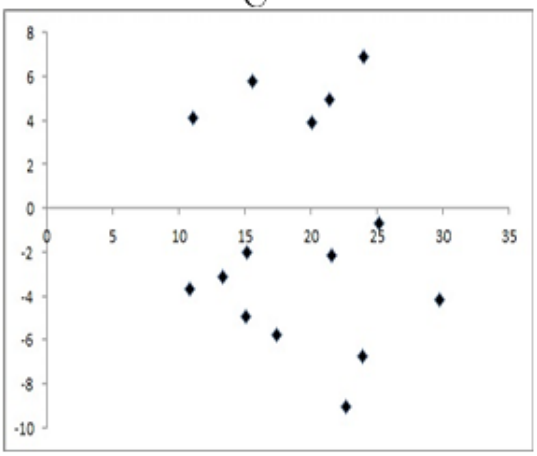
B



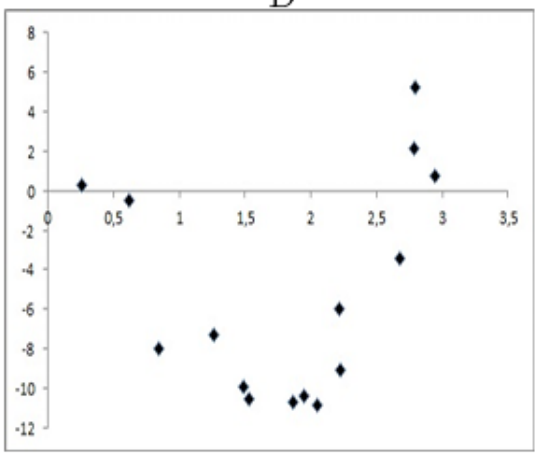
Il Grafico A evidenzia una dipendenza lineare positiva: all'aumentare di X aumenta anche Y; la Figura B, al contrario, evidenzia una dipendenza lineare negativa tra le due variabili: all'aumentare di una diminuisce l'altra.

I seguenti grafici a dispersione invece mostrano relazioni di tipo non lineare tra le variabili oggetto di studio.

C



D



Il Grafico C è costituito da una nuvola di punti informe e questo denota assenza di dipendenza (assenza di legame) tra le due variabili; il grafico D invece evidenzia una dipendenza curvilinea tra X e Y.

Il secondo strumento utile per studiare la dipendenza lineare tra due fenomeni è il coefficiente di correlazione (lineare), che in un certo senso quantifica le informazioni contenute in un diagramma a dispersione. Il coefficiente di correlazione lineare (di Bravais-Pearson), denotato con la lettera greca ρ, misura l'intensità della relazione lineare tra X e Y ed è dato dalla seguente formula:

$$\rho = \frac{\mathrm{Cov}(X;Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} = \frac{\tilde{S}_{XY}}{\tilde{S}_X \cdot \tilde{S}_Y}.$$

Si può dimostrare che $\boxed{-1 \leq \rho \leq 1}$. In particolare,

- $\rho = 1$ se e solo se i dati si allineano perfettamente lungo una retta inclinata positivamente.

  Più ρ assume valori prossimi a 1, più significa che esiste una relazione di tipo monotono crescente tra le variabili, come illustrato nel Grafico A.

- ρ = -1se e solo se i dati si allineano perfettamente lungo una retta inclinata negativamente.
  Più ρ assume valori prossimi a-1, più significa che esiste una relazione di tipo monotono decrescente tra le variabili, come illustrato nel Grafico B.

- ρ = 0 indica assenza di legame lineare tra X e Y e ciò può avvenire in due diverse situazioni: quando non sussiste alcun legame tra le due variabili come nel caso illustrato dal Grafico C e quando, pur essendoci un legame funzionale tra X e Y, questo non è di tipo lineare (ossia di tipo monotono crescente o decrescente). Quando ρ= 0 le variabili X e Y si dicono incorrelate. I dati riportati nel Grafici C e D sono esempi in cui $\rho \cong 0$.

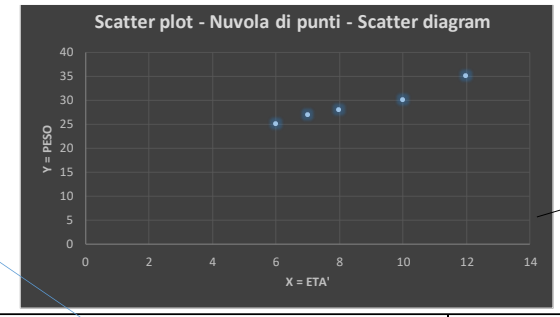
- ρ = 0 se e solo se la covarianza tra le variabili è nulla.

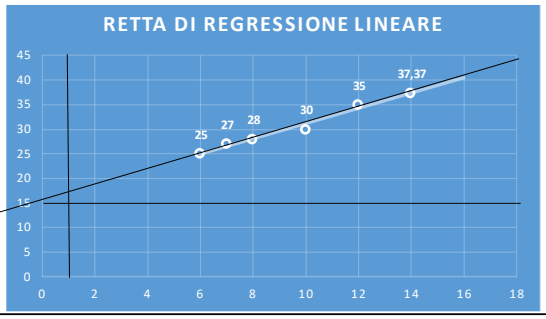
- Il segno di ρ coincide con il segno della covarianza tra X e Y. Il segno positivo indica che le due variabili aumentano o diminuiscono assieme, mentre il segno negativo indica che all'aumentare di una l'altra diminuisce.

Si ribadisce ancora una volta che se ρ è pari a zero (o prossimo a zero) significa che l'intensità del legame lineare è nulla (o bassissima) ma tra le due variabili potrebbe esserci una relazione (anche esatta) di altro tipo; il grafico a dispersione ci aiuta a comprendere se siamo in un caso di indipendenza (assenza di qualsiasi legame), come illustrato nel Grafico C o di dipendenza di tipo curvilineo come mostrato dalla Figura D.

ESERCIZIO SVOLTO

X ETA'	Y PESO
6	25
7	27
8	28
10	30
12	35
14	37,37





-1,5 -1 -0,6 0 0,8 1 1,5

Xmedia = (6+7+8+10+12)/5 = 43/5 = 8,6
Ymedia = 29
$S^2x = [(6-8.6)^2+(7-8.6)^2+(8-8.6)^2+(10-8.6)^2+(12-8.6)^2]/5 = 23.2/5 = 4.64$ (varianza di x)
Sx = 2,15 (DEVIANZAx = 23,2) (DEVIAZIONE STANDARD)
$S^2y = [(25-29)^2 + (27-29)^2 + (28-29)^2 + (30-29)^2 + (35-29)^2]/5 = 58/5 = 11.6$
Sy = 3,40
CODxy = ∑(X-Xmedia)(Y-Ymedia)
CODxy = [(6-8.6)(25-29)] + [(7-8.6)(27-29)] + [(8-8.6)(28-29)] +[(10-8.6)(30-29)] + [(12-8.6)(35-29)] → (-2.6)(-4) + (-1.6)(-2) + (-0.6)(-1) + (1.4)(1) + (3.4)(6) → +10.4 + 3.2 + 0.6 + 1.4 + 20.4 = 36
COVxy = 36/5 = 7.2
r = 7,2/(2,15x3,4) = 7,2/7,31 = 0,98 correlazione lineare positiva alta 0,32 -0,58

Y = A + BX (Y = $B_0 + B_1X$) A = 15,67 B = 1,55 **Y = 15,67 + 1,55X**

X=14 Y=?

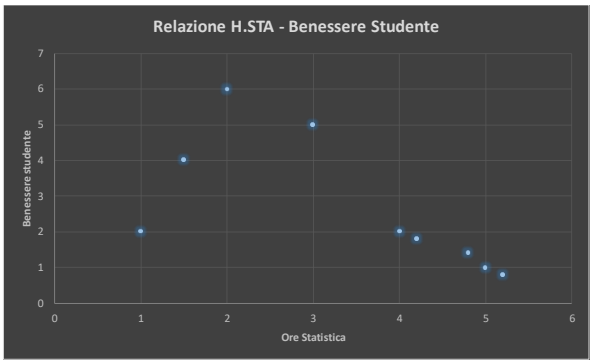
B = CODxy/DEVx ⟶ 36/23,2 = 1,55
A = Ymedia - BXmedia ⟶ 29 - 1,55 x 8,6 = 15,67

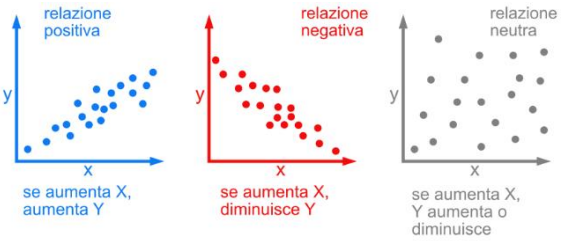
Y = 15,67 + 1,55X
Y = 15,67 + 1,55 x 14 = 37,37

BONTA' DI ADATTAMENTO O INDICE DI DETERMINAZIONE = $r^2 = 0,98^2 = 0,96$ ⟶ $0 < r^2 < 1$ 96%
INDICE DI ALIENAZIONE = $1 - r^2 = 1 - 0,96 = 0,04 = 4\%$

Nuvola Di Dispersione/Scatter Plot/Scatter Diagram

H STA	BENESS
1	2
1,5	4
2	6
3	5
4	2
4,2	1,8
4,8	1,4
5	1
5,2	0,8





Formula Del Coefficiente Di Correlazione Lineare Di Pearson

$$r_{xy} = \frac{\frac{\sum (X - \overline{X})(Y - \overline{Y})}{N}}{\sigma_X \sigma_Y}$$

CODEVIANZA
COD_{xy}

COVARIANZA
COV_{xy}

$-\sigma\sigma < COV_{xy} < +\sigma\sigma$ → $\pm\sigma\sigma$
coefficiente simmetrico

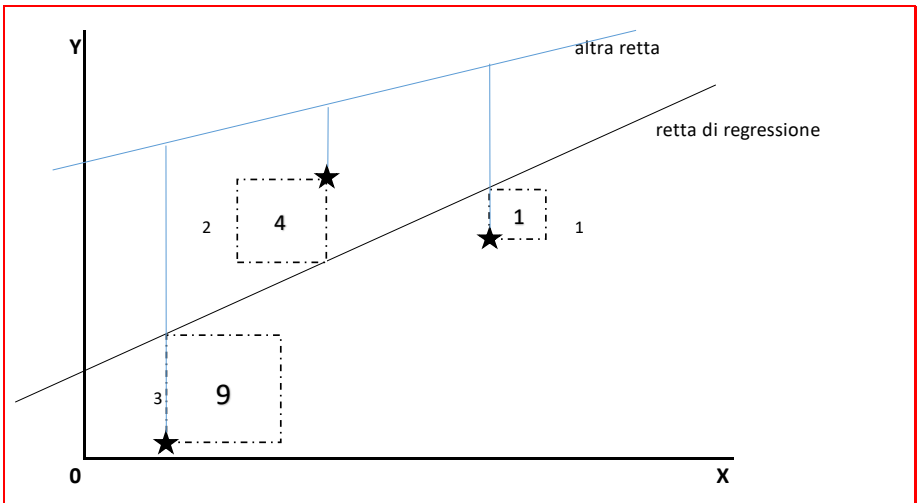
PRODOTTO DELLE DEVIAZIONI STANDARD

Correlazione

COEFFICIENTE DI CORRELAZIONE

- $-1 < r < +1$
- r=0 indica assenza di *correlazione*
- se r >0 le due variabili *covariano*
- se r <0 le due variabili *controvariano*
- r=1 o r=-1 esiste una relazione matematica tra le due variabili

Regola Dei Minimi Quadrati



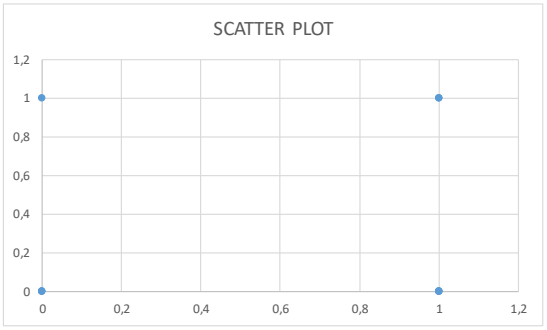
Metodo dei minimi quadrati: la retta di regressione minimizza la sommatoria dei quadrati costruiti sui residui ove per residuo si intende la distanza del nostro punto dalla stessa retta.

Esempio Esercizio

	1	2	3	4	5	6	7	8	9	10
X	0	1	1	1	0	0	1	1	1	0
Y	0	0	0	1	0	1	1	0	0	0

		X		
		0	**1**	**Ri**
Y	**0**	3	4	7
	1	1	2	3
	Ci	4	6	**10**

(0-0,6)(0-0,3)X3 + (1-0,6)(0-0,3)4....



Xmedia = 0,6
Ymedia = 0,3
Sx = 0,49
Sy = 0,46
CODxy = (0 - 0,6)(0 - 0,3)3 + (1 - 0,6)(0 - 0,3)4 + (0 - 0,6)(1 - 0,3) + (1 - 0,6)(1 - 0,3)2 = 0,2
COVxy = 0,02
r = 0,088

Capitolo VIII – Gini

L'indice di eterogeneità di Gini

Tale è un indice di eterogeneità **per variabili qualitative**. Esso offre una misura della eterogeneità (omogeneità) di una distribuzione statistica a partire dai valori delle frequenze relative associate alle k modalità di una generica variabile X (si ricordi che quest'indice è usato nella statistica descrittiva univariata, dove si considera una sola variabile). Ciò vuol dire che se i dati sono distribuiti in modo eterogeneo su tutte le k modalità di X (cioè, se le modalità hanno numerosità simili o, nel caso di massima eterogeneità, uguali), l'indice di Gini è elevato, viceversa, in caso di distribuzione di frequenza omogenea l'indice sarà (percentualmente) piuttosto basso. L'indice di Gini è così definito:

$$I = 1 - \sum_{i=1}^{k} f_i^2$$

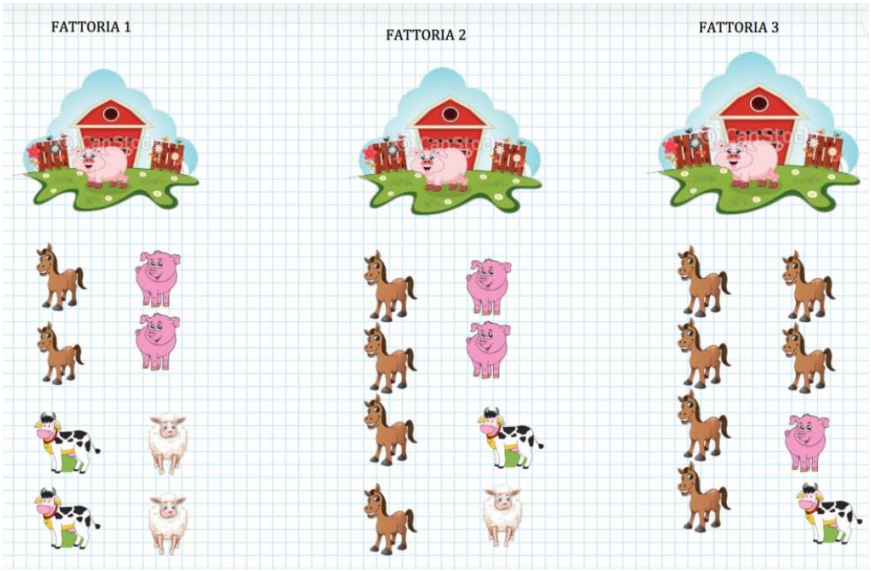
Per essere significativo, in generale, un indice deve essere relativizzato onde fornire una misura che sia confrontabile con altre al fine di stabilire se il valore trovato è più o meno elevato. Avendo due distribuzioni di n dati sulle quali si è calcolato l'indice di Gini, per avere informazioni sul grado di maggiore o minore elevatezza dell'eterogeneità devo opportunamente relativizzare l'indice, ottenendo così l'indice normalizzato. L'indice viene normalizzato dividendolo per il suo massimo se il minimo è uguale a 0 o dividendo l'indice meno il minimo dell'indice per il suo massimo meno il minimo se il minimo è diverso da 0. In questo modo si ottiene l'indice normalizzato (con valori tra 0 e 100%. Si calcolano perciò il minimo ed il massimo a partire da distribuzioni minimanti e massimanti.

In caso di **minima eterogeneità** (massima omogeneità) i dati sono tutti distribuiti su una modalità che ha frequenza pari al 100%, mentre le altre modalità hanno frequenza 0. In caso di **massima eterogeneità**, tutte le modalità hanno uguale frequenza.

Dividendo l'indice per il **massimo**, si ottiene l'**indice normalizzato** (percentuale):

$$I_N = \frac{I}{\frac{k-1}{k}} = \frac{k}{k-1} I.$$

L’ESEMPIO DELLE TRE FATTORIE



La maggior parte di voi sosterrà che quella più eterogenea è la prima in quanto presenta tutte e quattro le specie nella stessa misura.
In quanto tale è perciò quella che presenta una maggiore equi-distribuzione e quindi anche una più evidente diversità di genere.
Certamente la meno eterogenea è l’ultima fattoria che presenta meno delle quattro specie e una distribuzione più disomogenea.
Nell’ultima fattoria la specie “dominante” sono i cavalli che contano ben 6 esemplari, mentre di mucche e maialini ne troviamo solo uno.

Ma come possiamo fare a quantizzare all’interno della statistica questo fenomeno?

Calcolo dell’indice di Gini e normalizzazione dello stesso. Un indice normalizzato ha come **valore minimo 0** e come valore massimo 1. Interpretare questo valore è molto semplice.
Quando l’indice normalizzato assume valore zero diciamo che nella popolazione c’è **minima eterogeneità**, il che significa massima omogeneità. Detto in altre parole questo si verificherebbe quando abbiamo una fattoria che ha solo cavalli.
Quando l’l’indice di Gini normalizzato assume valore massimo, ovvero uno, la popolazione presenta la **massima eterogeneità**. In questo caso avremo che tutte le modalità presentano la stessa frequenza, cioè sono presenti tutte in egual misura.

fi	Fatt.1	Fatt.2	Fatt.3
Cavalli	2	4	6
Maiali	2	2	1
Mucche	2	1	1
Pecore	2	1	0
Totale	8	8	8

Fi	Fatt.1	Fatt.2	Fatt.3
Cavalli	0,25	0,5	0,75
Maiali	0,25	0,25	0,125
Mucche	0,25	0,125	0,125
Pecore	0,25	0,125	0
Totale	1	1	1

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

$G_1 = 1 - (4 * 0{,}25^2) = 0{,}75$

$G_2 = 1 - (0{,}5^2 + 0{,}25^2 + 2 * 0{,}125^2) = 0{,}65625$

$G_3 = 1 - (0{,}75^2 + 2 * 0{,}125^2) = 0{,}40625$

La prima fattoria è la più eterogenea

In realtà questo risultato ci può andare bene fin tanto che le popolazioni presentano lo stesso numero di unità statistiche e lo stesso numero di modalità.
In tutte e tre le fattorie c'erano 8 animali e li abbiamo suddivisi in quattro specie.
Per un risultato migliore e più attendibile abbiamo bisogno di normalizzare questo indice.

Fi	Fatt.1	Fatt.2	Fatt.3
Cavalli	0,25	0,5	0,75
Maiali	0,25	0,25	0,125
Mucche	0,25	0,125	0,125
Pecore	0,25	0,125	0
Totale	1	1	1

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

$G_1 = 1 - (4 * 0{,}25^2) = 0{,}75$

$G_2 = 1 - (0{,}5^2 + 0{,}25^2 + 2 * 0{,}125^2) = 0{,}65625$

$G_3 = 1 - (0{,}75^2 + 2 * 0{,}125^2) = 0{,}40625$

La prima fattoria è la più eterogenea

$$G_N = G * \frac{k}{k-1}$$

$G_{N1} = 0{,}75 * \frac{4}{3} = 1$

$G_{N2} = 0{,}65625 * \frac{4}{3} = 0{,}875$

$G_{N3} = 0{,}40625 * \frac{4}{3} = 0{,}542$

Come era lecito attendersi nella prima fattoria abbiamo trovato la massima eterogeneità possibile, con l'indice normalizzato pari al suo valore massimo 1.

L’indice di concentrazione di Gini

Tale è un indicatore che offre una misura della concentrazione di **variabili quantitative trasferibili** ossia è una misura della diseguaglianza di una distribuzione. È spesso usato come indice di concentrazione per misurare la diseguaglianza nella distribuzione del reddito o anche della ricchezza. È un numero compreso tra 0 ed 1. Valori bassi del coefficiente indicano una distribuzione abbastanza omogenea, con il valore 0 che corrisponde alla pura equidistribuzione, ad esempio la situazione in cui tutti percepiscono esattamente lo stesso reddito; valori alti del coefficiente indicano una distribuzione più diseguale, con il valore 1 che corrisponde alla massima concentrazione, ovvero la situazione dove una persona percepisca tutto il reddito del paese mentre tutti gli altri hanno un reddito nullo. A volte l’indice di Gini viene moltiplicato per cento, diventando così un valore tra 0 e 100, più facile da visualizzare graficamente e da comprendere nei suoi trend di crescita o di decrescita.

Indice di Gini	Concentrazione
$0 \leq G < 0{,}25$	Bassa
$0{,}25 \leq G < 0{,}5$	Medio-bassa
$0{,}5 \leq G < 0{,}75$	Medio-alta
$0{,}75 \leq G \leq 1$	Alta